RICE UNIVERSITY

Study of Influenza Virus Hemagglutinin

By

Jun Shen

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Jianpeng Ma, Professor, Chair
Department of Bioengineering, RICE
Biochemistry & Molecular Biology, BCM

Yizhi Jane Tao, Assistant Professor
Department of Biochemistry & Cell Biology
RICE University

Bruce R. Johnson, Distinguished Faculty Fellow
Department of Chemistry, RICE University
Executive Director, Rice Quantum Institute

HOUSTON, TEXAS

November 2009

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

Study of Influenza Virus Hemagglutinin

by
Jun Shen

The suddenly global outbreak of 2009 H1N1 Flu reminds human being the danger and severity of influenza virus. The gene of the new strain come from five different influenza viruses: North American swine influenza, North American avian influenza, human influenza, and two swine influenza viruses typically found in Eurasia. The recombination of gene in virus evolvement makes it more and more important to understand the evolutionary characteristics of influenza.

Hemagglutinin (HA), embedded on the surface of influenza virus, is one of two virally-coded integral envelope proteins of the virus. The three primary functions of hemagglutinin (HA) include receptor binding, membrane fusion, and antigenic variation. Study of HA structure and its evolutionary mechanism is crucial to fully understand influenza virus.

In the first chapter, a study of diversifying selective pressure on influenze B virus hemagglutinin was reported. All the positively selected sites were located in the four epitopes (120-loop, 150-loop, 160-loop and 190-helix) of HA, and all of them have been identified in previous studies. This supports a predominant role of antibody selection in

HA evolution.

In the second chapter, we studied positive selection analysis of 2009 H1N1 Flu. Among a subgroup of human A(H1N1) HAs between 1918~2008, we found strong diversifying (positive) selection at HA_1 156 and 190. We also analyzed the evolutionary trends at HA_1 190 and 225 that are critical determinants for receptor-binding specificity of A(H1N1) HA. Additional analysis of directional selection was also employed for H1N1 gene data.

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor, **Dr. Qinghua Wang**, for being the best teacher I met in my student life. She taught me every step of biology experiment, and instructed me for details with extremely patience. Dr. Wang has super memory and sharp mind, works with full enthusiasm for research. All the computation and analysis was based on her comprehensive and profound understanding of the structure and function of Influenza virus. I've been very lucky to work under directions of Dr. Wang, otherwise I may have difficulty to graduate.

I would like to thank my committee member **Dr. Bruce R. Johnson** and **Dr. Yizhi Jane Tao**, for providing me invaluable guidance and insight.

Six years ago, I was desperately looking for a decent Ph.D. program, and **Dr. Jianpeng Ma** recruited me from MSU. I would like to give my special thanks to Dr. Ma. I am so proud to be an OWL! I am so proud of being Dr. Ma's student!

I would like to thank previous and current members of the Ma group: Dr Xia Tian, Dr. Yifei Kong, Dr. Yinghao Wu, Dr. Mingzhi Chen, Mingyang Lu, Brian Kirk, Xiaorui Chen, Fengyun Ni, Cheng Zhang, for their meaningful discussion and friendship. Thanks **Dr. Junhua Pan** for his help of X-ray data diffraction and crystallization robot.

Thanks to my wife and my parents. They always encourage me when I get frustrated, support me when I am weak.

Contents

CHAPTER One

Diversifying Selective Pressure on Influenza B Virus Hemagglutinin

1.1 Abstract

Influenza B virus hemagglutinin (HA) is a major surface glycoprotein with frequent amino-acid substitutions. However, the roles of antibody selection in the amino-acid substitutions of HA were still poorly understood. In order to gain insights into this important issue, an analysis was conducted on a total of 271 HA_1 sequences of influenza B virus strains isolated during 1940~2007. In this analysis, PAML (Phylogenetic Analysis by Maximum Likelihood) package was used to detect the existence of positive selection and to identify positively selected sites on HA_1. Strikingly, all the positively selected sites were located in the four major epitopes (120-loop, 150-loop, 160-loop and 190-helix) of HA identified in previous studies, thus supporting a predominant role of antibody selection in HA evolution. Of particular significance is the involvement of the 120-loop in positive selection, which may become increasingly important in future field isolates. Despite the absence of different subtypes, influenza B virus HA continued to evolve into new sublineages, within which the four major epitopes were targeted selectively in positive selection. Thus, any newly emerging strains need to be placed in the context of their evolutionary history in order to understand and predict their epidemic potential.

Key Words: Positive selection/Antigenic drift/Molecular evolution/Antibody selection

1.2 Introduction

Ever since the isolation of the first influenza B virus strain B/Lee/40 [Krystal et al., 1982], influenza B virus has remained a serious health problem, contributing to the seasonal "flu" epidemics each year. As a major glycoprotein on the surface of influenza B virus, HA undergoes constant amino-acid substitutions. The HA protein of current circulating influenza B virus strains belongs to one of the two major phylogenetic lineages: B/Victoria/2/87 (B/VI)-like and B/Yamagata /16/88 (B/YM)-like [Kanegae et al., 1990; Rota et al., 1990; Shaw et al., 2002].

Over the last 68 years, a large number of amino-acid substitutions on influenza B virus HA were observed in field isolates, in monoclonal-antibody escape mutants and in egg-adapted variants [Berton et al., 1984; Berton and Webster, 1985; Bootman and Robertson, 1988; Hovanec and Air, 1984; Krystal et al., 1982; Krystal et al., 1983; Lubeck et al., 1980; Rota et al., 1992; Rota et al., 1990; Verhoeyen et al., 1983; Webster and Berton, 1981]. However, it was unclear which of these substitutions the results of positive selection were, and what were the roles of antibody selection in the molecular evolution of influenza B virus HA. In this context, **positive selection** is defined as a significant excess of amino-acid altering substitutions over silent substitutions in nucleotide sequences, since, if completely random, only 24% of nucleotide substitutions would cause changes in the encoded amino acids [Air et al., 1990].

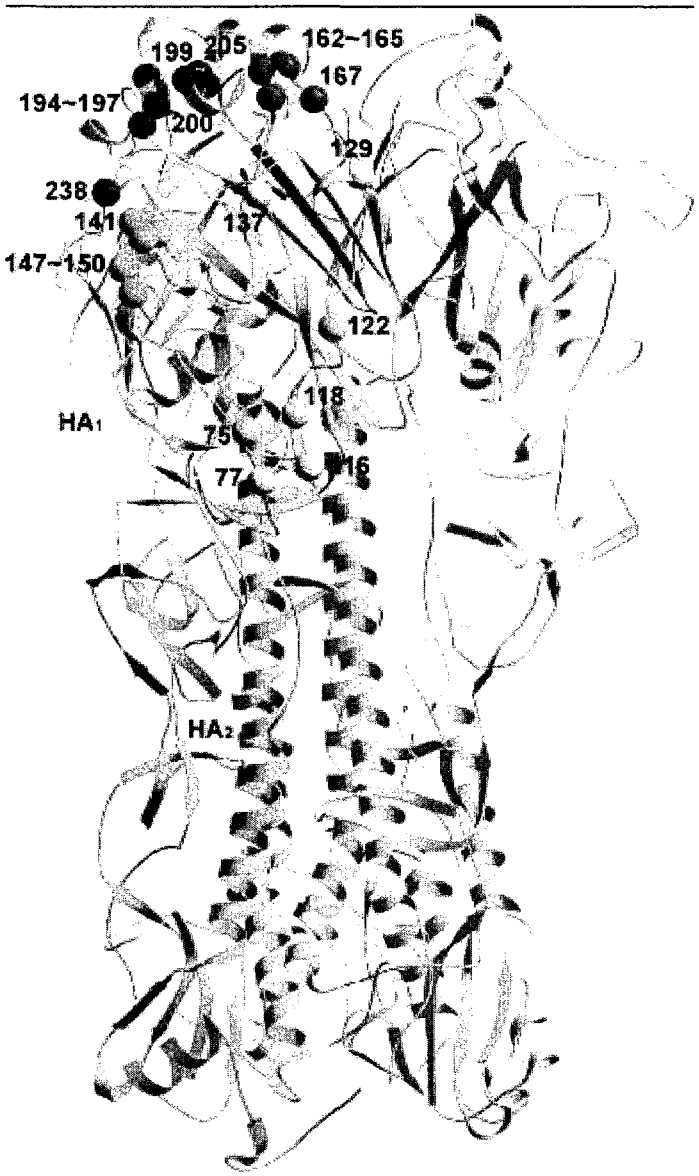


Fig 1.1: Major epitopes of influenza B virus HA. The trimeric HA is shown with one monomer highlighted in color: Pink for HA1 and yellow for HA2. Mutations in four regions, the 120-loop (cyan), 150-loop (green), 160-loop (blue), and 190-helix (red), have been found to cause antigenicity variation.

There was previously a sequence analysis on 49 HA_1 sequences of recent influenza B virus isolates, which identified HA_1 75, 197, and 199 (B/HongKong/8/73 HA numbering of 75, 194, and 196, respectively) to be under positive evolutionary selection [Pechirra et

al., 2005]. However, since it did not separate the B/YM-lineage and B/VI-lineage strains, this study might have failed to identify those amino-acid positions that were selected positively only in one but not the other lineage [Yang et al., 2000]. This was particularly problematic for the 150-loop and 160-loop (**Fig.1.1**), which had become specific for B/YM-like and B/VI-like strains, respectively [Nakagawa et al., 2001a; Nakagawa et al., 2001b; Nakagawa et al., 2003; Nakagawa et al., 2005; Wang et al., 2008]. Most recently, a larger-scale analysis that used 214 HA_1 sequences of influenza B virus strains has been published [Chen and Holmes, 2008]. Although it separated B/YM- and B/VI-lineage strains, the evolution of these lineages into distinct sublineages was not taken into account, which limited the accuracy of the positively selected sites derived therein [Chen and Holmes, 2008].

PAML is a package of programs that analyze DNA and protein sequences using maximum likelihood [Yang, 2007]. Using the program CODEML in PAML, the nonsynonymous (amino-acid altering substitutions)/synonymous (silent substitutions) rate ratio (ω) for each codon is calculated as an important indicator of selection pressure at the protein level: an $\omega > 1$ indicates positive selection [Yang, 2007; Yang et al., 2000]. Bayes Empirical Bayes analysis then calculates the posterior probability that each site belongs to a particular site class. Sites with high posterior probability of belonging to the site class of $\omega > 1$ are inferred to be under positive selection [Yang, 2007; Yang et al., 2000].

In order to gain insights into the amino-acid positions on influenza B virus HA that are truly under positive selective pressure, here a total of 271 HA_1 sequences of influenza

B virus strains isolated between 1940~2007 were analyzed. Based on the phylogenetic analysis, these HA_1 sequences were divided into three major groups: early strains (1940~1970), B/YM-like lineage (1972~2005) and B/VI-like lineage (1975~2007). The B/YM-lineage was further divided into four sublineages, and the B/VI lineage into two sublineages (**Fig. 1.4**). These seven groups were analyzed by using CODEML in PAML version 4 [Yang, 2007; Yang et al., 2000]. The identified positively selected sites were located predominantly on the four major antigenic epitopes on HA_1: the 120-loop (HA_1 116~137), the 150-loop (HA_1 141~150), the 160-loop (HA_1 162~167), the 190-helix (HA_1 194~202), and their respective surrounding regions [Wang et al., 2008] (**Fig. 1.5**), suggesting the important roles of antibody selection in molecular evolution of influenza B virus HA.

1.3 Materials and Methods

Phylogenetic analysis

a) Sequences preparation

This study focused on the first 340 amino acid residues of mature HA_1 (1~1020 nucleotides excluding those corresponding to the signal peptide). A total of 271 HA_1 sequences of influenza B virus strains isolated between 1940~2007 were used in the study. These sequences were selected to sample all the years in which influenza B viruses were active, and special cares were taken to avoid stains with high similarity and isolated in the same regions. All the sequences were obtained from the Influenza Sequence Database (Los Alamos National Laboratory, Los Alamos, NM, USA www.flu.lanl.gov) [Macken et al., 2001]. The name of the strain was composed of two part, strain type plus serial number and year, for example:

NewYork 2 – 90

NewYork: The location where the virus was collected

2: Serial number, there may exit another strain NewYork3-90.

90: The virus strain was at 1990.

Fig 1.2: Meaning of the strain name.

All the adopted sequences are saved as fasta format, which are compatible with DNA star for further analysis.

b) Alignment of Sequence and tree drawing

The CLUSTAL W method [Thompson et al., 1994] with the MEGALIGN program of

DNASTAR package (www.dnastar.com) was employed for sequencing alignment and phylogenetic analysis. Among all the available methods of sequence alignment, clustal W method provide the best sensitivity for the alignment of divergent protein sequences. Although most of the sequences used in this work share high similarity, the sequence shift in alignment is foremost step for calculation of substitution rate ratio. For Clustal W method, **PAM250** residue weight table was adopted.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Table 1.1: Residue Weights of Clustal W method.

Here is an example of the alignment result of the early strains, one of the seven sublineages we studied, is listed as **fig 1.3** as the following:

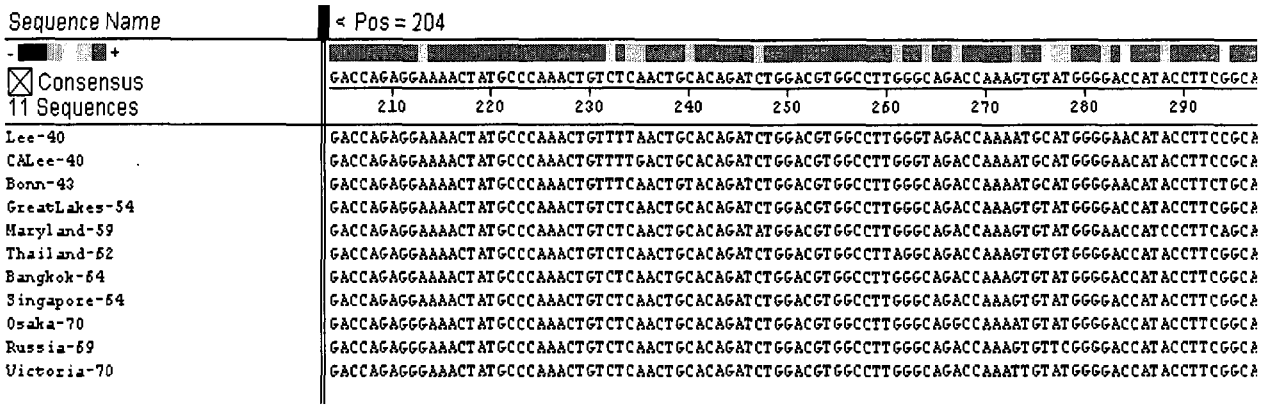


Fig. 1.3: Alignment of early strain.

In these position marked by red, all the strains possess same sequence, no variance. Our calculation focused on these non red areas, which are positions under evolution pressure, either synonymous (silent) or nonsynonymous(amino acid altering) substitution.

All the evolution tree pictures in this paper were also made by DNASTAR, and the tree format is rooted tree.

c) Analysis of selective pressure and setting of PAML

PAML is open source software for phylogenetic analysis of DNA or protein sequence. It analyses all the sequence through comparison of nested statistical models.

For this analysis, the CODEML program in PAML was used to calculate the codon-substitution models for heterogeneous selection pressure at amino-acid positions [Yang, 1997; Yang, 2007; Yang et al., 2000; Yang et al., 2005]. The models used in this study were M0, M1a, M2a, M7 and M8.

Model	NSsites	#p	Parameters
M0 (one ratio)	0	1	ω
M1a (neutral)	1	2	p_0 ($p_1 = 1 - p_0$), $\omega_0 < 1$, $\omega_1 = 1$
M2a (selection)	2	4	p_0, p_1 ($p_2 = 1 - p_0 - p_1$), $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 > 1$
M3 (discrete)	3	5	p_0, p_1 ($p_2 = 1 - p_0 - p_1$) ω_0, ω_1, ω_2
M7 (beta)	7	2	p, q
M8 (beta&ω)	8	4	p_0 ($p_1 = 1 - p_0$), p, q, $\omega_s > 1$

Table 1.2. Parameters in the site models, #p is the number of free parameters in the ω distribution.

M1a (nearly neutral) and M7 (beta) were null models that did not support $\omega > 1$. In contrast, the alternative models M2a (positive selection) and M8 (beta and ω), compared to M1a and M7 respectively, each had an additional class that allowed $\omega > 1$. Likelihood ratio tests (LRT) comparing M2a versus M1a and M8 versus M7 provided test for the existence of positive selection. In LRT, twice the log likelihood difference, $2\Delta l = 2(l_1 - l_0)$, was compared with a χ^2 distribution to test whether the null model was to be rejected, where ℓ_0 and ℓ_1 were the log likelihood for the alternative

model and the null model, respectively. In addition, empirical Bayes analysis was employed to calculate the posterior probability that each site belonged to a particular site class. Sites with high posterior probability of belonging to the site class of $\omega > 1$ were inferred to be under positive selection. It was shown that Bayes Empirical Bayes, which assigned a prior to the model parameters [Deely and Lindley, 1981], worked well for both small and large datasets [Yang et al., 2005]. Since some of the subgroups used in this study were small, the results from Bayes Empirical Bayes analysis were used throughout this study. To account for the insertions and deletions in influenza B virus HA_1, the numbering of influenza B/HongKong/8/73 HA was used as a reference for all sequences [Wang et al., 2008].

The key control file for CODEML program, Codeml.ctl contains all important parameter for the calculation,

seqfile = FluB.NUC * *sequence data file name*

treefile = Flub..DND * *tree structure file name*

outfile = mlc * *main result file name*

noisy = 3 * *0,1,2,3,9: how much rubbish on the screen*

verbose = 0 * *1: detailed output, 0: concise output*

runmode = 0 * *0: user tree; 1: semi-automatic; 2: automatic*

* *3: StepwiseAddition; (4,5):PerturbationNNI; -2: pairwise*

CodonFreq = 2 * *0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table*

clock = 0 * *0: no clock, unrooted tree, 1: clock, rooted tree*

aaDist = 0 * *0:equal, +:geometric; -:linear, {1-5:G1974,Miyata,c,p,v}*

model = 0

NSsites = 0 8 * *Mode0 amd Model 8 will be caculated in this run*

* *0:one w; 1:NearlyNeutral; 2:PositiveSelection; 3:discrete;*

* *4:freqs;5:gamma;6:2gamma;7:beta;*

* *8:beta&w;9:betaγ10:3normal*

icode = 0 * *0:standard genetic code; 1:mammalian mt; 2-10:see below*

Mgene = 0 * *0:rates, 1:separate; 2:pi, 3:kappa, 4:all*

fix_kappa = 0 * *1: kappa fixed, 0: kappa to be estimated*

kappa = .3 * *initial or fixed kappa*

fix_omega = 0 * *1: omega or omega_1 fixed, 0: estimate*

omega = 1.3 * *initial or fixed omega, for codons or codon-based AAs*

fix_alpha = 0 * 0: estimate gamma shape parameter; 1: fix it at alpha

alpha = 0.5 * initial or fixed alpha, 0:infinity (constant)

ncatG = 10 * *# of categories in the dG or AdG models of rates*

getSE = 0 * *0: don't want them, 1: want S.E.s of estimates*

RateAncestor = 0 * *(0,1,2): rates (alpha>0) or ancestral states (1 or 2)*

Small_Diff = .45e-6

cleandata = 1 * *remove sites with ambiguity data (1:yes, 0:no)*

fix_blength = 0 * *0: ignore, -1: random, 1: initial, 2: fixed*

Several pivotal input of CODEML worth to mention and discuss here:

Model: set equal to 0, that assume all the branches possess the same ratio ω.

Alpha: refers to the parameter α in gamma distribution for variable substitution rates across sites.

Clock: Without the clock (clock = 0), unrooted trees should be used, such as ((1,2),3,4) or (1,2,(3,4)).

1.4 Results

Phylogenetic relationship of influenza B virus HA

According to phylogenetic analysis, the 271 HA_1 sequences were divided into three groups: early strains isolated between 1940~1970 (I), B/YM-like lineage since 1972 (II) and B/VI-like lineage since 1975 (III). The B/YM-lineage (II) was divided further into four (II-*i* ~ II-*iv*) sublineages (**Fig. 1.4**), among which (II-*i* ~ II-*iii*) sublineages had been described in a previous study [Nerome et al., 1998], whilst the (II-*iv*) sublineage was described here for the first time. The B/VI-lineage (III) was divided further into an earlier sublineage (III-*i*) and a more recent sublineage (III-*ii*) (**Fig. 1.4**). This large-scale phylogenetic analysis uncovered that the divergence of influenza B virus HA into B/YM- and B/VI-lineages can be dated back to early 1970s, which is much earlier than previously thought [Kanegae et al., 1990; Matsuzaki et al., 2004; McCullers et al., 2004; Rota et al., 1990] and agrees well with a just-published study [Chen et al., 2007].

Positive selection on influenza B virus HA

To detect positively selected sites in HA_1 sequence of influenza B virus strains between 1940~2007, the analysis using CODEML in PAML was performed individually on the seven subgroups (I, II-*i* ~ II-*iv*, and III-*i* ~ *ii*) (**Fig. 1.4**). In all but two cases, the LRT statistics ($2\Delta l$) for M2a versus M1a and M8 versus M7 were much larger than the critical value of $\chi^2_{1\%} = 6.63$ with degree of freedom (d.f.) set to 1 (**Table 1.3, 1.4**). Thus the LRT tests supported the existence of positive selection on influenza B virus HA. The sites with greater than 50% posterior probability to be under positive selective pressure in models M2a and M8, obtained from Bayes Empirical Bayes analysis [Yang et

al., 2005], were listed in **Table 1.5**. In general, M2a identified fewer sites under positive selection than M8 did. Nevertheless, the sites identified in M2a were those of the highest posterior probability in M8 (**Table 1.5**). In contrast, those identified only in M8 but not in M2a were generally of low posterior probability. To be more conservative, most of our discussion was focused on the sites that were identified in M8 model with greater than 95% posterior probability to be under positive selection. This cutoff limits the false-positive rate to 5~6% or lower [Yang et al., 2005]. It is important to emphasize that those of high posterior probability to be under positive selection were not necessarily those of the highest mutation rates. Different from influenza A virus HA [Bush et al., 1999; Yang et al., 2000], a much smaller number of sites on influenza B virus HA were subject to positive selection for antigenic drift, consistent with earlier studies [Air et al., 1990; Chen and Holmes, 2008].

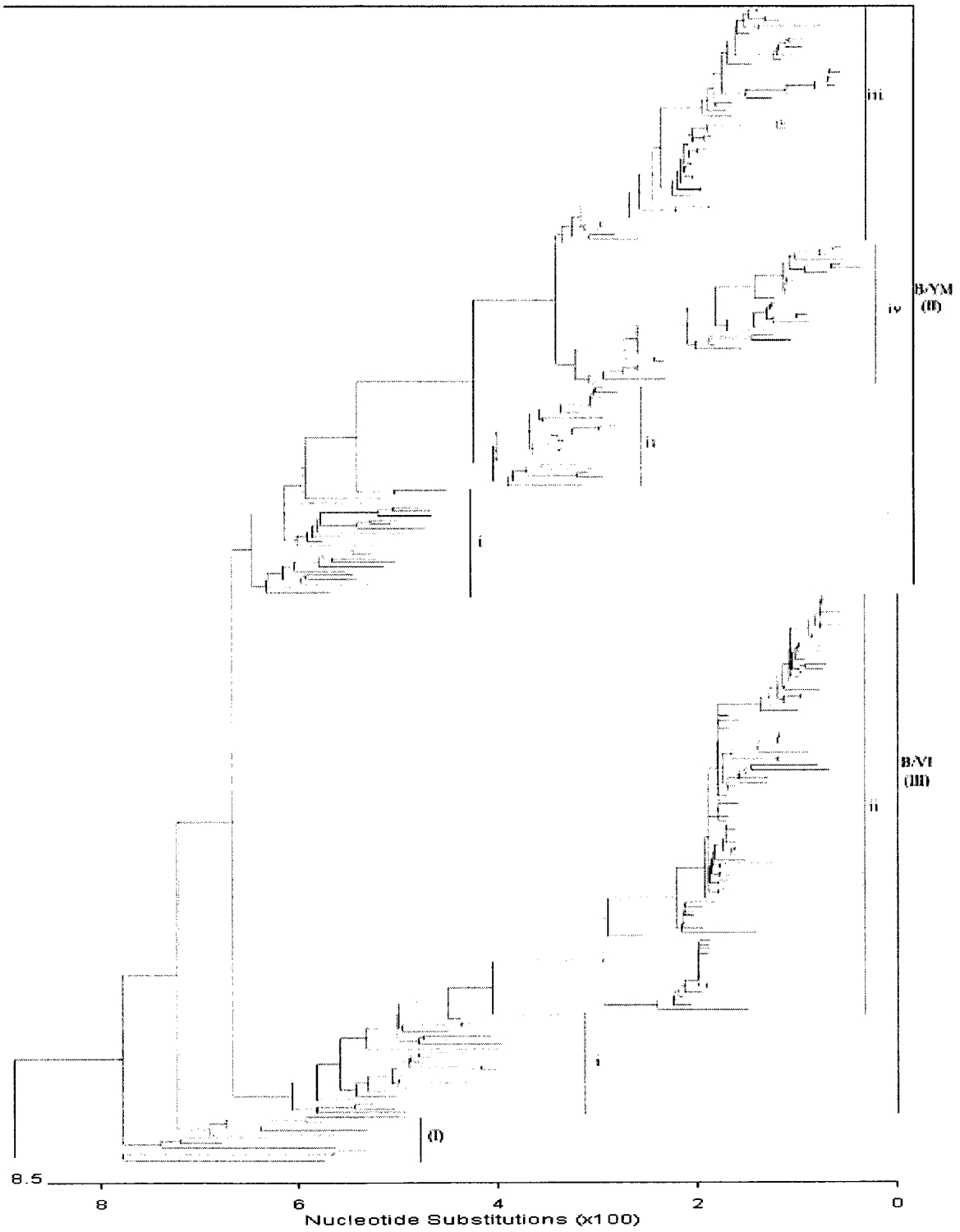


Fig. 1.4: Phylogenetic relationship of 271 HA1 sequences used in this study. For all sequences, the nucleotide sequences between 1 and 1,020, corresponding to residues HA1 1–340, were used. The phylogenetic tree was drawn using the program Megalign from DNASTAR package (www.dnastar.com).

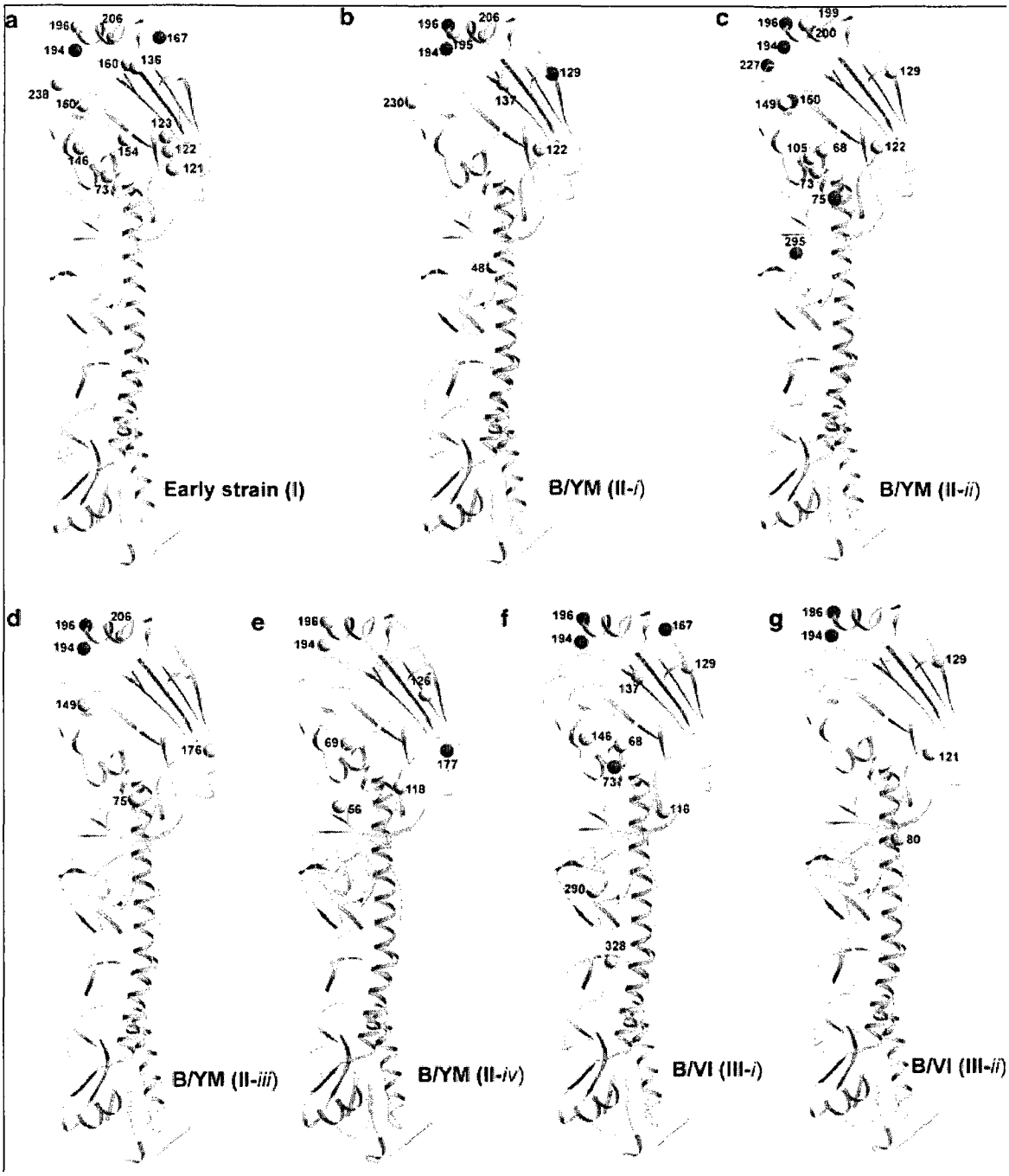


Fig 1.5: Sites with posterior probabilities of greater than50% to be under positive selection in the M8 models for the seven subgroups of influenza B virus HA, in the order of early strain (I) (a), B/YM-lineage (II-i–II-iv) (b–e) and B/VI-lineage (III-i and III-ii) (f, g). Each site is shown as a ball centered at its Ca atom in the structure (Protein Data Bank code 3BT6) [Wang et al., 2008]. Sites with greater than 95% posterior probability to be under positive selection are shown in dark color and the rest are in light color. The structure of one monomer of HA is in the same orientation as the monomer shown in color in Figure 1.1.

Table1.3: The values of log-likelihood (ℓ), d_N/d_S, and parameter estimates in the analysis of the HA_1 subunit of influenza B virus strains circulating between 1940~2007.

Model	ℓ	d_N/d_S	Parameters estimates
Early strain (I) 1940~1970 (11 strains)			
M0 (one-ratio)	-2343.24	0.271	ω=0.271
M1a (nearly neutral)	-2317.53	0.258	p_0=0.744 (p_1=0.256), ω_0=0.002 (ω_1=1)
M2a (positive selection)	-2314.85	0.297	p_0=0.744, p_1=0.251 (p_2=0.005), ω_0=0.005 (ω_1=1), ω_2=7.990
M7 (beta)	-2317.89	0.236	p=0.016, q=0.051
M8 (beta&ω)	-2315.01	0.287	p_0=0.994 (p_1=0.006), p=0.017, q=0.051, ω_s=7.428
B/YM-lineage (II-*i*) 1972~1984 (25 strains)			
M0 (one-ratio)	-2501.04	0.373	ω=0.373
M1a (nearly neutral)	-2461.96	0.262	p_0=0.755 (p_1=0.245), ω_0=0.022 (ω_1=1)
M2a (positive selection)	-2439.17	0.404	p_0=0.729, p_1=0.261 (p_2=0.011), ω_0=0.020 (ω_1=1), ω_2=11.904
M7 (beta)	-2462.35	0.300	p=0.005, q=0.012
M8 (beta&ω)	-2439.28	0.426	p_0=0.989 (p_1=0.011), p=0.017, q=0.042, ω_s=12.282
B/YM-lineage (II-*ii*) 1987~1996 (24 strains)			
M0 (one-ratio)	-2032.98	0.311	ω=0.311
M1a (nearly neutral)	-2002.40	0.188	p_0=0.812 (p_1=0.188), ω_0=0 (ω_1=1)
M2a (positive selection)	-1995.15	0.318	p_0=0.826, p_1=0.136 (p_2=0.038), ω_0=0 (ω_1=1), ω_2=4.779
M7 (beta)	-2002.45	0.200	p=0.005, q=0.020
M8 (beta&ω)	-1995.32	0.313	p_0=0.953 (p_1=0.047), p=0.012, q=0.076, ω_s=4.237
B/YM-lineage (II-*iii*) 1991~2002 (56 strains)			
M0 (one-ratio)	-2211.56	0.200	ω=0.200
M1a (nearly neutral)	-2196.38	0.168	p_0=0.901 (p_1=0.099), ω_0=0.077 (ω_1=1)
M2a (positive selection)	-2190.27	0.213	p_0=0.936, p_1=0.056 (p_2=0.009), ω_0=0.105 (ω_1=1), ω_2=6.802
M7 (beta)	-2197.92	0.184	p=0.098, q=0.437
M8 (beta&ω)	-2190.51	0.212	p_0=0.991 (p_1=0.009), p=0.431, q=2.312, ω_s=6.740
B/YM-lineage (II-*iv*) 1994~2005 (33 strains)			
M0 (one-ratio)	-2144.23	0.220	ω=0.220
M1a (nearly neutral)	-2127.98	0.211	p_0=0.789 (p_1=0.211), ω_0=0 (ω_1=1)
M2a (positive selection)	-2127.40	0.242	p_0=0.847, p_1=0.022 (p_2=0.131), ω_0=0.021 (ω_1=1), ω_2=1.537
M7 (beta)	-2128.03	0.200	p=0.005, q=0.020
M8 (beta&ω)	-2127.40	0.242	p_0=0.868 (p_1=0.132), p=0.052, q=1.023, ω_s=1.560
B/VI-lineage (III-*i*) 1975~1993 (24 strains)			
M0 (one-ratio)	-2530.27	0.336	ω=0.336
M1a (nearly neutral)	-2492.37	0.251	p_0=0.749 (p_1=0.251), ω_0=0 (ω_1=1)
M2a (positive selection)	-2477.70	0.351	p_0=0.736, p_1=0.254 (p_2=0.010), ω_0=0.005 (ω_1=1), ω_2=9.700
M7 (beta)	-2492.81	0.222	p=0.009, q=0.029
M8 (beta&ω)	-2477.89	0.354	p_0=0.983 (p_1=0.017), p=0.016, q=0.050, ω_s=7.267
B/VI-lineage (III-*ii*) 1996~2007 (98 strains)			
M0 (one-ratio)	-2924.64	0.299	ω=0.299
M1a (nearly neutral)	-2899.56	0.266	p_0=0.805 (p_1=0.195), ω_0=0.088 (ω_1=1)
M2a (positive selection)	-2887.74	0.320	p_0=0.796, p_1=0.199 (p_2=0.005), ω_0=0.094 (ω_1=1), ω_2=9.871
M7 (beta)	-2899.97	0.266	p=0.156, q=0.430
M8 (beta&ω)	-2887.00	0.309	p_0=0.994 (p_1=0.006), p=0.244, q=0.698, ω_s=8.534

Table 1.4: Likelihood ratio tests (LRT) between M2a versus M1a and M8 versus M7 for the seven subgroups of HA_1 subunit of influenza B virus strains circulating between 1940~2007

LRT	$2\Delta\ell^*$
Early strain (I) 1940~1970 (11 strains)	
M2a – M1a	**5.36**
M8 – M7	**5.76**
B/YM-lineage (II-*i*) 1972~1984 (25 strains)	
M2a – M1a	**44.44**
M8 – M7	**46.14**
B/YM-lineage (II-*ii*) 1987~1996 (24 strains)	
M2a – M1a	**14.50**
M8 – M7	**14.26**
B/YM-lineage (II-*iii*) 1991~2002 (56 strains)	
M2a – M1a	**12.22**
M8 – M7	**14.82**
B/YM-lineage (II-*iv*) 1994~2005 (33 strains)	
M2a – M1a	**1.16**
M8 – M7	**1.26**
B/VI-lineage (III-*i*) 1975~1993 (24 strains)	
M2a – M1a	**29.34**
M8 – M7	**29.84**
B/VI-lineage (III-*ii*) 1996~2007 (98 strains)	
M2a – M1a	**23.64**
M8 – M7	**25.94**

* In LRT tests, the values of $2\Delta\ell$ were compared with the critical values of χ^2 distribution (9.21 and 5.99 for $\chi^2_{1\%}$ and $\chi^2_{5\%}$, respectively, with d.f.=2) [Yang, 1997; Yang, 2007; Yang et al., 2000; Yang et al., 2005]. Significantly larger values of $2\Delta\ell$ over those of χ^2 distributions led to the rejection of the null models M1a and M7.

Table 1.5 Sites with higher than 50% posterior probabilities of being under positive selective pressure for the HA_1 subunit of influenza B virus strains circulating between 1940~2007

Model	Positively selected sites[1]
	Early strain (I) 1940~1970 (11 strains)
M2a (positive selectio	73, 150, 167*, 194**, 196, 238
M8 (beta&ω)	73*, 121, 122, 123, 136, 146, 150*, 154*, 160, 167***, 194***, 196**, 206, 238*
	B/YM-lineage (II-*i*) 1972~1984 (25 strains)
M2a (positive selectio	122, 129***, 194***, 196***, 206
M8 (beta&ω)	48, 122**, 129***, 137, 194***, 195, 196***, 206**, 230
	B/YM-lineage (II-*ii*) 1987~1996 (24 strains)
M2a (positive selectio	68*, 75**, 122, 129, 150**, 194***, 196***, 200, 227**, 295**
M8 (beta&ω)	68**, 73, 75***, 105, 122*, 129**, 149, 150***, 194***, 196***, 199, 200, 227***, 295***
	B/YM-lineage (II-*iii*) 1991~2002 (56 strains)
M2a (positive selectio	176, 194***, 196***
M8 (beta&ω)	75, 149, 176*, 194***, 196***, 206
	B/YM-lineage (II-*iv*) 1994~2005 (33 strains)
M2a (positive selectio	69*, 177**, 194, 196
M8 (beta&ω)	56, 69**, 118, 126, 177***, 194**, 196**
	B/VI-lineage (III-*i*) 1975~1993 (24 strains)
M2a (positive selectio	73*, 116, 167**, 194***, 196***, 290, 328
M8 (beta&ω)	68, 73***, 116**, 129, 137, 146, 167***, 194***, 196***, 290**, 328**
	B/VI-lineage (III-*ii*) 1996~2007 (98 strains)
M2a (positive selectio	194***, 196***
M8 (beta&ω)	80, 121*, 129, 194***, 196***

[1]Positively selected sites from Bayes Empirical Bayes analysis [Yang et al., 2005].
*Posterior probability of positive selective pressure is between 75~84%.
**Posterior probability of positive selective pressure is between 85~94%.
***Posterior probability of positive selective pressure is higher than 95%.

Early strain (I) (1940~1970).

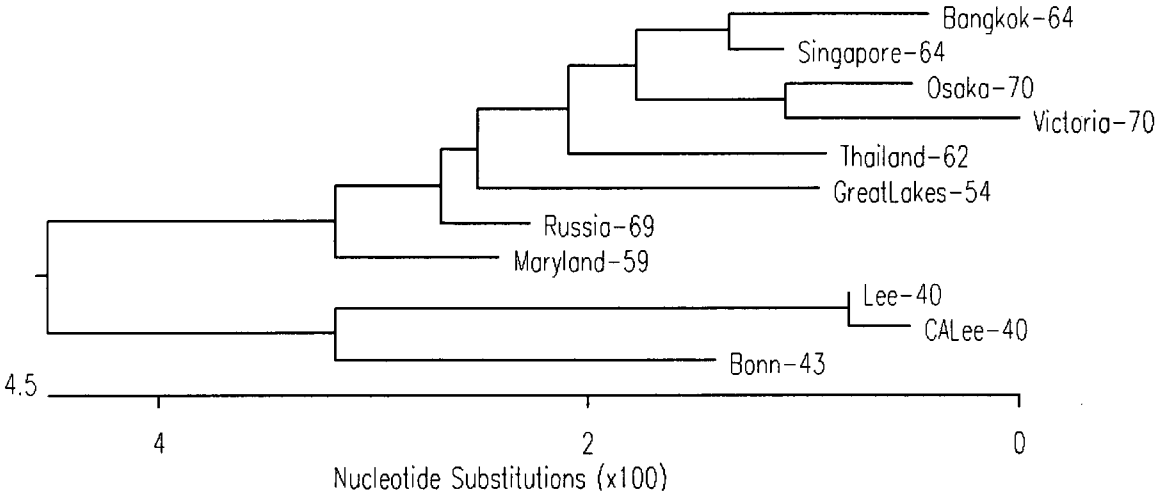


Fig. 1.6: Early strain lineage (I).

Among the 271 HA_1 sequences analyzed in this study, a total of 11 sequences over a time span of 31 years belong to this group (**Fig. 1.4, Fig. 1.6**). To limit the uncertainties related to the relatively small number of samples in this group, the results from Bayes Empirical Bayes analysis were used throughout this study [Anisimova et al., 2002; Yang et al., 2005]. In LRT tests, the values of $2\Delta l$ were 5.36 for M2a versus M1a, and 5.76 for M8 versus M7 (**Table 1.4**). These values were larger than the critical value of $\chi^2_{5\%}$ = 3.84, but smaller than $\chi^2_{1\%}$ = 6.63 with d.f. = 1 [Yang, 1997; Yang, 2007; Yang et al., 2000; Yang et al., 2005]. The M2a model suggested ~0.5% sites to be under positive selection with ω_2=7.990 (**Table 1.3**). Similarly, the M8 model suggested ~0.6% sites to be under positive selection with ω_s=7.428. The M2a model identified a total of six sites to be under positive selective pressure (>50% posterior probability) (**Table 1.5**). The M8 model identified 14 sites of being under positive selective pressure (>50% posterior probability) (**Fig. 1.6**). Among them, two sites were of greater than 95% posterior

probability to be under positive selection: HA_1 167 (95%) on the 160-loop and 194 (99%) on the 190-helix.

B/YM-like lineage (II).

A total of 138 HA_1 sequences in this analysis belong to B/YM-like lineage. It was further divided into four sublineages, II-*i* (25 sequences), II-*ii* (24 sequences), II-*iii* (56 sequences) and II-*iv* (33 sequences) (**Fig. 1.4**).

Early strain sublineage (II-*i*) (1972~1984).

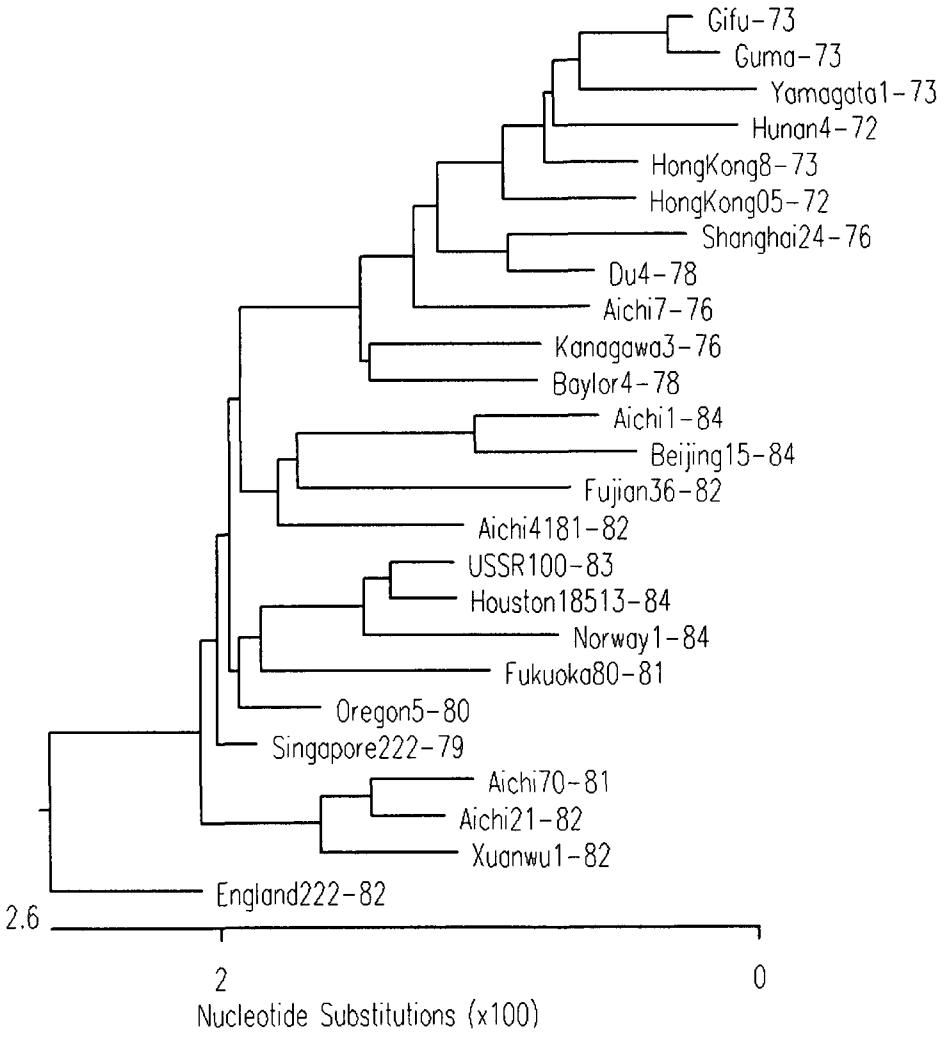


Fig. 1.7: B/YM-lineage II-*i*(1972~1984)

These early strains of B/YM-lineage spanned a period of 13 years (Fig. 1.7). The values of M2a versus M1a and M8 versus M7 were much greater than = 6.63 with d.f. = 1 (**Table 1.3, 1.4**), resulting in the rejection of the null models M1a and M7. Both M2a and M8 models suggested ~1.1% sites to be under strong positive selection with large values (**Table 1.3**). The M2a model identified a total of five sites to be under positive selection (>50% posterior probability) (Table 1.5), three of which were of greater than 95% posterior probability: HA1 129 (97%) on the 120-loop, 194 (100%) and 196 (100%) on the 190-helix. These three sites were again with >95% posterior probability in the M8 model: HA1 129 (99%), 194 (100%) and 196 (100%) (**Fig. 1.7**).

Sublineage (II-*ii*) (1987~1996).

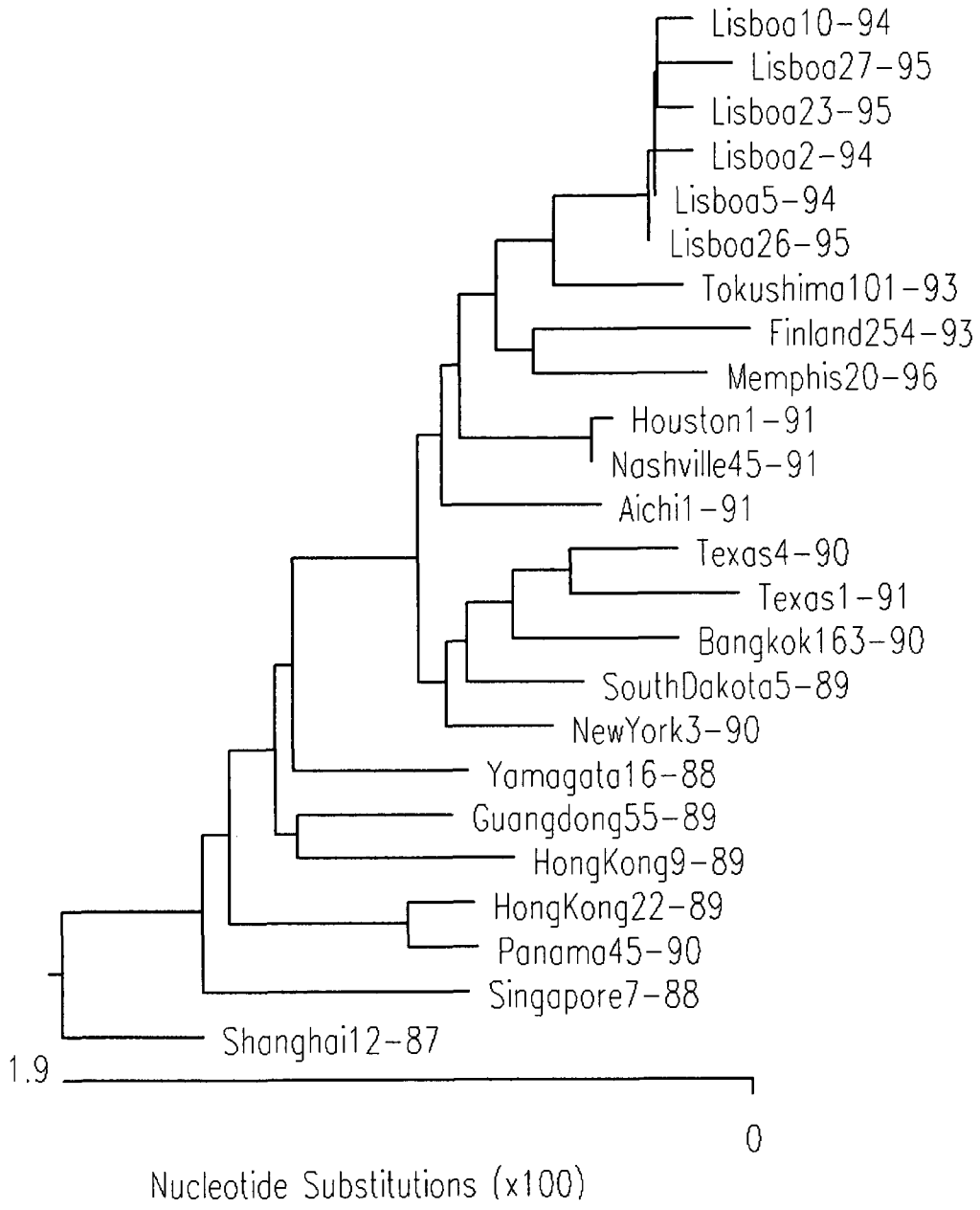


Fig. 1.8: B/YM-lineage II-*ii*

The B/YM-lineage strains in this group covered a 10-year period (**Fig. 1.8**). In LRT tests, the $2\Delta l$ values of M2a versus M1a and M8 versus M7 provided strong support for the existence of positive selection (**Table 1.3, 1.4**). Both M2a and M8

models suggested ~4% sites to be under positive selection with $\omega \approx 4$ (**Table 1.3**). The M2a model identified two sites with higher than 95% posterior probability of being positively selected (**Table 1.5**): HA_1 194 (99%) and 196 (97%) on the 190-helix. The M8 model identified a total of six sites with greater than 95% posterior probability of being positively selected (**Table 1.5**): HA_1 75 (97%) and 295 (98%) on the 120-loop, 150 (96%) on the 150-loop, 194 (100%), 196 (99%) and 227 (97%) on the 190-helix (**Fig. 1.8**). It is noteworthy that HA_1 150 on the 150-loop was inferred to be under positive selection with very high confidence, in excellent agreement with previous conclusions that the 150-loop is an important epitope for B/YM-lineage [Nakagawa et al., 2001a; Nakagawa et al., 2003].

Sublineage (II-*iii*) (1991~2002). This sublineage of B/YM-like strains covered a 12-year period (**Fig. 1.9**). The LRT tests led to the rejection of the null models M1a and M7 (**Table 1.3, 1.4**). Both M2a and M8 models suggested ~0.9% sites to be under positive selection with $\omega \approx 7$ (**Table 1.3, 1.4**). The M2a model revealed three sites of being under positive selection, including HA_1 194 (97%) and 196 (99%) on the 190-helix (**Table 1.5**). These two sites were of 99% and 100% posterior probability of positive selection in the M8 model (**Table 1.5 and Fig. 1.9**).

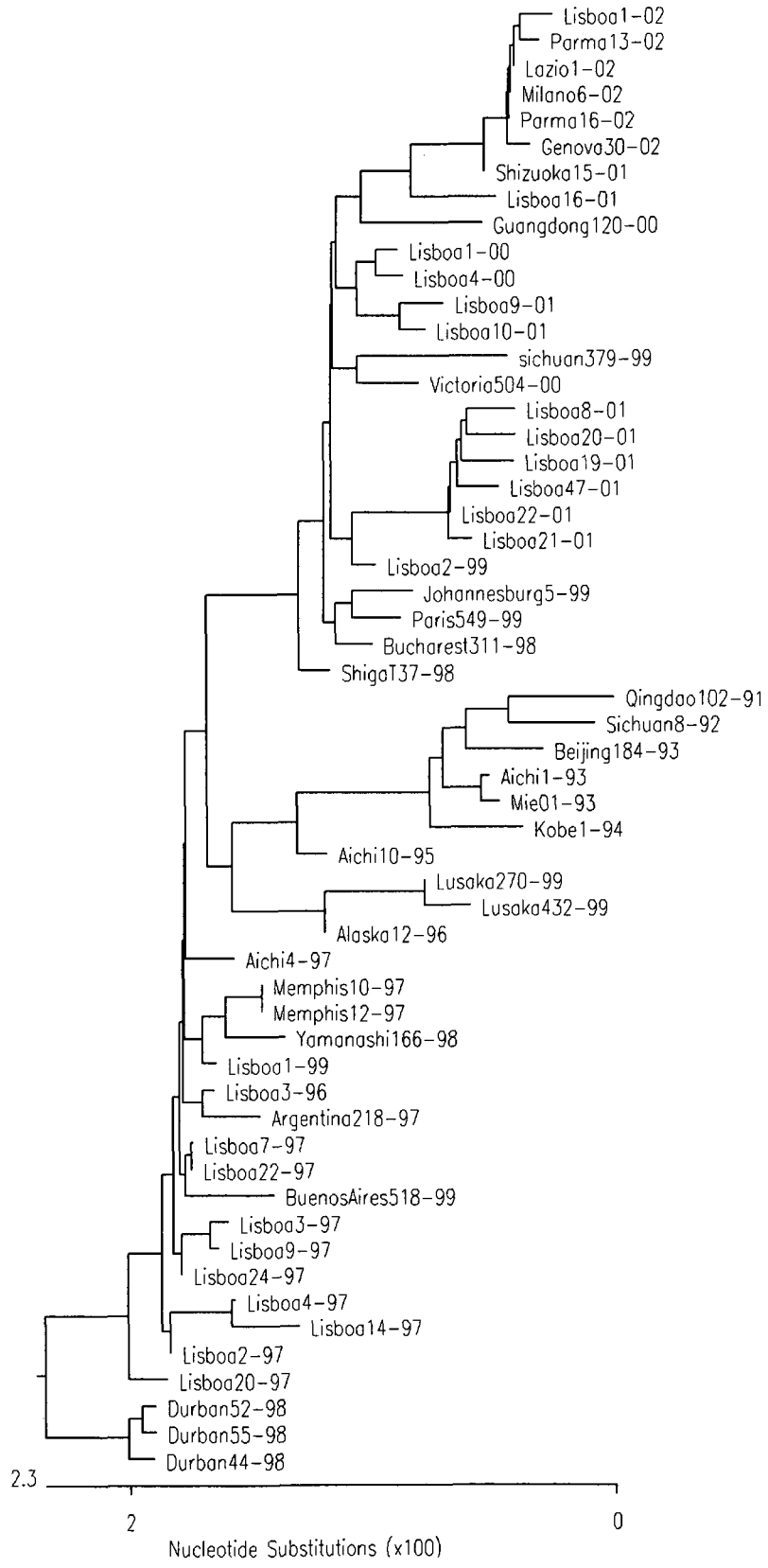


Fig. 1.9: B/YM-lineage II-*iii*.

Sublineage (II-*iv*) (1994~2005).

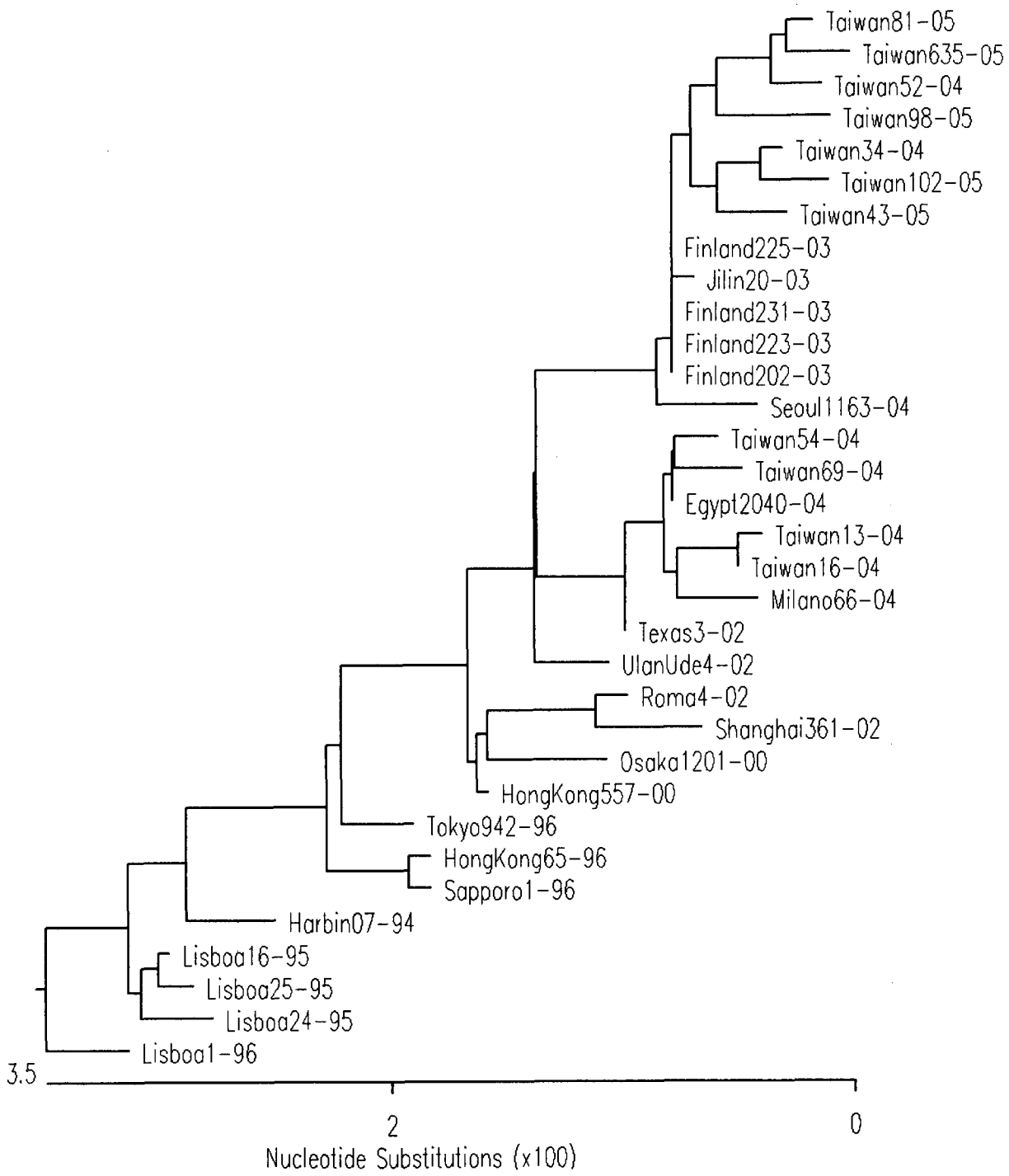


Fig. 1.10: B/YM-lineage II-*iv*.

This sublineage of B/YM-like strains contained some of the most recently circulating

strains of B/YM-lineage (**Fig. 1.10**). In sharp contrast to all other sublineages of B/YM-like strains and to all B/VI-like strains, the LRT statistics were $2\Delta l$ =1.16 and 1.26 for M2a versus M1a and M8 versus M7, respectively (**Table 1.4**), suggesting a low confidence for the existence of positive selection. In Bayes Empirical Bayes analysis, both M2a and M8 models suggested a relatively large percentage of sites (~13%) to be under very weak positive selection with ω_2=1.537 and ω_s=1.560, respectively (**Table 1.3**). The M2a model identified a total of four positively selected sites with >50% posterior probability (**Table 1.5**). In the M8 model, a total of seven sites were identified, with only one site, HA_1 177 (98%) on the 120-loop, with > 95% posterior probability (**Table 1.5 and Fig. 1.10**). This sublineage was the only group in which HA_1 194 and 196 are of lower than 95% probability to be under positive selection.

B/VI-like lineage (III). A total of 122 HA_1 sequences of influenza B virus strains belong to this lineage. They were grouped into two sublineages, early strains (III-*i*) containing 24 sequences and more recent strains (III-*ii*) containing 98 sequences (**Fig. 1.4**).

Early strain sublineage (III-*i*) (1975~1993). These early strains of B/VI-lineage spanned a time period of 19 years and exhibited significant sequence differences from the recent circulating B/VI-like strains (III-*ii*) (**Fig.2, Fig. 1.11**). The LRT statistics supported the rejection of the null models (M1a and M7) and strongly supporting the presence of positive selection (**Table 1.3, 1.4**). The M2a model suggested 1.0% sites to be under positive selection with ω_2=9.700 (**Table 1.3**). Similarly, the M8 model suggested 1.7% sites to be under positive selection with ω_s=7.267. The

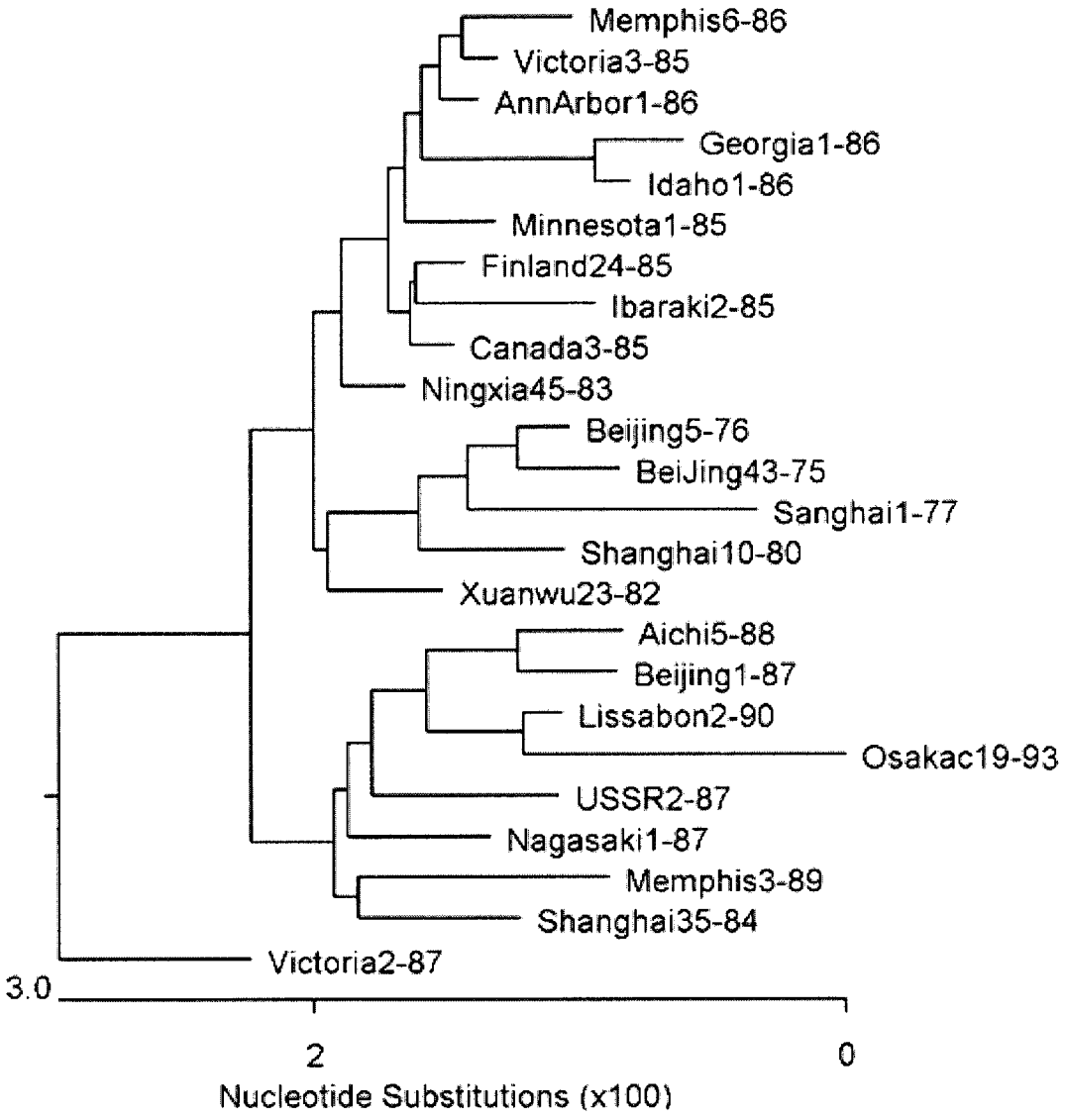


Fig. 1.11: B/VI-lineage III-*i*.

M2a model identified seven sites to be under positive selective pressure (**Table 1.5**), with HA_1 194 (100%) and 196 (100%) on the 190-helix of higher than 95% posterior probability. The M8 model revealed a total of 11 positively selected sites, among which four sites were of greater than 95% posterior probability. They were HA_1 73 (95%) (120-loop), 167 (97%) (160-loop), 194 (100%) and 196 (100%) (190-helix) (**Table 1.5,**

Fig. 1.11). It is important to note that among these four sites with the highest posterior probability, one site, HA_1 167, is on the 160-loop, while none is located on the 150-loop. In sharp contrast, B/YM-like (II-*ii*) sublineage, which circulated in an overlapping time period and contained the same number of sequences, had HA_1 150 on the 150-loop to be under positive selection. These observations further supported earlier conclusions that the 160-loop epitope is specific for the B/VI-lineage strains [Nakagawa et al., 2001b; Nakagawa et al., 2005] while the 150-loop epitope is specific for the B/YM-lineage strains [Nakagawa et al., 2001a; Nakagawa et al., 2003].

Recent strain sublineage (III-*ii*) (1996~2007).

The more recent isolates of B/VI-lineage strains remained to be a single group over the time period of 12 years (**Fig. 1.12**). The LRT statistics supported strongly the presence of positive selection (**Table 1.3, 1.4**). The M2a model suggested ~0.5% sites to be under positive selection with ω_2=9.871 (**Table 1.3**). Similarly, the M8 model suggested ~0.6% sites to be under positive selection with ω_s=8.534. HA_1 194 had a posterior probability of 95% and 99% to be under positive selection in M2a and M8 models, respectively, while HA_1 196 has a 100% posterior probability in both M2a and M8 models (**Table 1.5 and Fig. 1.12**). Compared to the earlier B/VI-like (III-*i*) sublineage, one noticeable difference is that the 160-loop was no longer under positive selection in these recent strains (III-*ii*). Rather, positive selection was focused on the 190-helix.

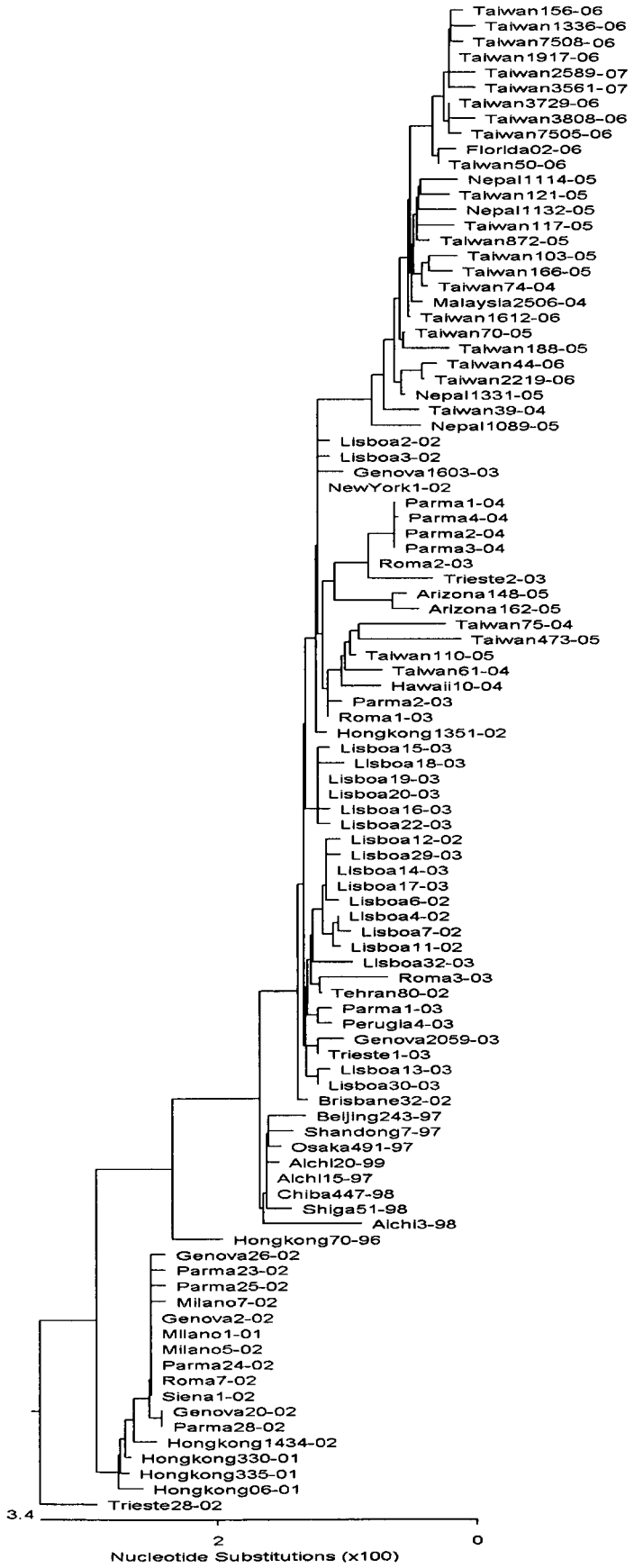


Fig. 1.12:

B/VI-lineage III-ii.

1.5 Discussion and Conclusions

Roles of antibody selection in the evolution of influenza B virus HA

In previous studies, four major antigenic epitopes of influenza B virus HA, the 120-loop, the 150-loop, the 160-loop, and the 190-helix, were identified on the membrane-distal domain of HA_1 [Wang et al., 2008] (**Fig.1.1**). Strikingly, in this study, all the identified positively selected sites in the seven subgroups were located on these four major antigenic epitopes, supporting the important roles of antibody selection in the molecular evolution of influenza B virus HA.

The 150-loop is an important epitope on HA. Antigenic properties were altered for influenza B virus with mutations on this loop in laboratory-selected escape mutants [Berton et al., 1984; Berton and Webster, 1985; Nakagawa et al., 2003; Webster and Berton, 1981], field isolates [Abed et al., 2003; Nakagawa et al., 2001a] and egg-adapted variants [Lugovtsev et al., 2005; Lugovtsev et al., 2007; Oxford et al., 1991; Oxford et al., 1990]. In more recent influenza B virus isolates, the 150-loop region appeared to be the neutralizing epitope specific for B/YM-like strains [Nakagawa et al., 2001a; Nakagawa et al., 2003]. Consistent with that finding, HA_1 150 was under positive selection with 96% posterior probability in B/YM-like (II-*ii*) sublineage.

The 160-loop is the only region in influenza B virus HA where insertions, deletions and single amino-acid substitutions were detected in field isolates [McCullers et al., 1999; Nakagawa et al., 2005; Nerome et al., 1998] and mAb-escape mutants [Berton et al., 1984; Berton and Webster, 1985; Nakagawa et al., 2001a; Nakagawa et al., 2001b; Webster and Berton, 1981], as an effective way for influenza B virus to survive a long period of time without antigenic shifts as observed in influenza A virus [Nerome et al.,

1998]. In recent isolates, the 160-loop became specific for B/VI-like lineage [Nakagawa et al., 2001b; Nakagawa et al., 2005]. In agreement with this observation, HA_1 167 on the 160-loop was selected positively in early strains (I) and in B/VI-like (III-*i*) sublineage (**Table 1.5, Fig.1.5**), with 95% and 96% posterior probability, respectively.

The 190-helix, which forms part of the receptor-binding site (RBS) of influenza B virus HA, is inarguably one of the most important epitopes. The hot spot is at HA_1 194~196, a potential glycosylation site. Similar to influenza A virus HA [Caton et al., 1982; Daniels et al., 1983; Schulze, 1997; Skehel et al., 1984; Skehel and Wiley, 2000], influenza B virus HA also utilized the addition or removal of glycosylation as a mechanism for antigenic drift [Berton et al., 1984; Berton and Webster, 1985; Gambaryan et al., 1999; Ikonen et al., 2005; Muyanga et al., 2001; Nakagawa et al., 2000; Nakagawa et al., 2004; Oxford et al., 1991; Oxford et al., 1990; Robertson et al., 1990; Robertson et al., 1985; Saito et al., 2004; Schild et al., 1983; Wang et al., 2008]. In this current analysis, HA_1 194 and 196 were constantly identified to be under positive selective pressure, with greater than 99% probability in 11 out of 14 cumulative cases (combining both sites in seven groups), and over 85% in three other cases. HA_1 227 in sublineage (II-*ii*) was another positively selected site on 190-helix with high posterior probability (97%) (**Table 1.5, Fig. 1.5**).

Perhaps one of the most important observations from this study is positive selection of the 120-loop region. The 120-loop epitope was defined as HA_1 116~137 and its surrounding regions [Wang et al., 2008]. In this context of this article, we refer to all sites not adjoining the 150-loop, 160-loop or the 190-helix epitopes as the 120-loop region due to spatial proximity (**Fig.1.1**). Although the 120-loop region appeared to be

one of the most frequently mutated regions in field isolates [Verhoeyen et al., 1983], its role in antigenicity of influenza B virus HA was not recognized until most recently [Lugovtsev et al., 2007; Nakagawa et al., 2006; Wang et al., 2008]. One possibility for such a delay in recognition is that the 120-loop is proximal to the viral envelope membrane, making the access by antibodies more difficult, as observed for influenza A virus HA [Barbey-Martin et al., 2002; Bizebard et al., 2001; Bizebard et al., 1995; Fleury et al., 1999; Gigant et al., 2000; Knossow et al., 2002]. Thus, it is very important that this current study provided strong evidence for positive selection of the 120-loop region, further supporting its significance in antigenicity of influenza B virus HA.

Trends of positive selection on influenza B virus HA

The early strains (I) seemed to have rather even distribution of positive selective pressure on all four major epitopes, although the positive selection on the 160-loop and 190-helix appeared to be stronger and/or more prevailing (**Table 1.5, Fig. 1.6**). In contrast, the early strains of B/YM-lineage and B/VI-lineage, sublineages (II-*ii*) and (III-*i*) respectively, were sharply divided. HA_1 150 on the 150-loop in B/YM-like (II-*ii*) sublineage, and HA_1 167 on the 160-loop in B/VI-like (III-*i*) sublineage, were inferred to be under positive selection with high posterior probability (**Table 1.5**). These observations agreed very well with earlier studies in which the 150-loop and 160-loop were found to be specific epitopes for the B/YM- and B/VI-lineages, respectively [Nakagawa et al., 2001a; Nakagawa et al., 2001b; Nakagawa et al., 2003; Nakagawa et al., 2005]. However, despite large sequence differences, the recent B/YM-like (II-*iii*) sublineage and B/VI-like (III-*ii*) sublineage converged at focusing on the 190-helix for

antigenic drift (**Table 1.5 and Fig. 1.5**). Most strikingly, in the newest B/YM-like (II-*iv*) sublineage, a large number of sites were found to be under rather weak positive selection, and the only positively selected site identified with high confidence was HA_1 177 on the 120-loop. The new trends of positive selection among these most recent strains, in conjunction with results from other studies [Lugovtsev et al., 2007; Nakagawa et al., 2006], stress the increasingly important role of the 120-loop in antigenicity of influenza B virus HA.

Concluding remarks

This study reports a large-scale systematic analysis of diversifying positive selective pressure on HA of distinct lineages/sublineages of influenza B virus isolated in the past 68 years. The highlights of the results from this study are:

a). The number of positively selected sites in influenza B virus HA were much fewer than those of influenza A virus HA [Air et al., 1990];

b). Although it does not have subtypes as influenza A virus HA, influenza B virus HA did and continue to diverge into different sublineages. This was particularly true for B/YM-lineage, as exemplified by the newly emerging B/YM-like (II-*iv*) sublineage that had not been previously described.

c). The study revealed the predominant roles of antibody selection in the molecular evolution of influenza B virus HA.

d). Despite the differences among different lineages/sublineages, HA_1 194 and 196 were constantly under positive selective pressure in all but one cases.

e). The 120-loop was an important epitope under constant positive selection. It

may play an increasingly important role in antigenicity in future field isolates, as evidenced in the most recent B/YM-like (II-*iv*) sublineage (**Table 1.5**).

f). Each lineage/sublineage utilized their respective favorite sites in positive selection. Thus, for any newly emerging strains of influenza B virus, it is important to put them in the context of their evolutionary history in order to understand and appreciate their full epidemic potential.

1.6 References

1. Abed Y, Coulthart MB, Li Y, Boivin G. 2003. Evolution of surface and nonstructural-1 genes of influenza B viruses isolated in the Province of Quebec, Canada, during the 1998-2001 period. Virus Genes 27(2):125-135.
2. Air GM, Gibbs AJ, Laver WG, Webster RG. 1990. Evolutionary changes in influenza B are not primarily governed by antibody selection. Proc Natl Acad Sci U S A 87(10):3884-3888.
3. Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19(6):950-958.
4. Barbey-Martin C, Gigant B, Bizebard T, Calder LJ, Wharton SA, Skehel JJ, Knossow M. 2002. An antibody that prevents the hemagglutinin low pH fusogenic transition. Virology 294(1):70-74.
5. Berton MT, Naeve CW, Webster RG. 1984. Antigenic structure of the influenza B virus hemagglutinin: nucleotide sequence analysis of antigenic variants selected with monoclonal antibodies. J Virol 52(3):919-927.
6. Berton MT, Webster RG. 1985. The antigenic structure of the influenza B virus hemagglutinin: operational and topological mapping with monoclonal antibodies. Virology 143(2):583-594.
7. Bizebard T, Barbey-Martin C, Fleury D, Gigant B, Barrere B, Skehel JJ, Knossow M. 2001. Structural studies on viral escape from antibody neutralization. Curr Top Microbiol Immunol 260:55-64.
8. Bizebard T, Gigant B, Rigolet P, Rasmussen B, Diat O, Bosecke P, Wharton SA, Skehel JJ, Knossow M. 1995. Structure of influenza virus haemagglutinin

complexed with a neutralizing antibody. Nature 376(6535):92-94.

9. Bootman JS, Robertson JS. 1988. Sequence analysis of the hemagglutinin of B/Ann Arbor/1/86, an epidemiologically significant variant of influenza B virus. Virology 166(1):271-274.

10. Bush RM, Fitch WM, Bender CA, Cox NJ. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol 16(11):1457-1465.

11. Caton AJ, Brownlee GG, Yewdell JW, Gerhard W. 1982. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell 31(2 Pt 1):417-427.

12. Chen JM, Guo YJ, Wu KY, Guo JF, Wang M, Dong J, Zhang Y, Li Z, Shu YL. 2007. Exploration of the emergence of the Victoria lineage of influenza B virus. Arch Virol 152(2):415-422.

13. Chen R, Holmes EC. 2008. The Evolutionary Dynamics of Human Influenza B Virus. J Mol Evol.

14. Daniels RS, Douglas AR, Skehel JJ, Wiley DC. 1983. Analyses of the antigenicity of influenza haemagglutinin at the pH optimum for virus-mediated membrane fusion. J Gen Virol 64 (Pt 8):1657-1662.

15. Deely JJ, Lindley DV. 1981. Bayes empirical bayes. J Am Stat Assoc 76:833-841.

16. Fleury D, Barrere B, Bizebard T, Daniels RS, Skehel JJ, Knossow M. 1999. A complex of influenza hemagglutinin with a neutralizing antibody that binds outside the virus receptor binding site. Nat Struct Biol 6(6):530-534.

17. Gambaryan AS, Robertson JS, Matrosovich MN. 1999. Effects of egg-adaptation on

the receptor-binding properties of human influenza A and B viruses. Virology 258(2):232-239.

18. Gigant B, Barbey-Martin C, Bizebard T, Fleury D, Daniels R, Skehel JJ, Knossow M. 2000. A neutralizing antibody Fab-influenza haemagglutinin complex with an unprecedented 2:1 stoichiometry: characterization and crystallization. Acta Crystallogr D Biol Crystallogr 56 (Pt 8):1067-1069.

19. Hovanec DL, Air GM. 1984. Antigenic structure of the hemagglutinin of influenza virus B/Hong Kong/8/73 as determined from gene sequence analysis of variants selected with monoclonal antibodies. Virology 139(2):384-392.

20. Ikonen N, Pyhala R, Axelin T, Kleemola M, Korpela H. 2005. Reappearance of influenza B/Victoria/2/87-lineage viruses: epidemic activity, genetic diversity and vaccination efficacy in the Finnish Defence Forces. Epidemiol Infect 133(2):263-271.

21. Kanegae Y, Sugita S, Endo A, Ishida M, Senya S, Osako K, Nerome K, Oya A. 1990. Evolutionary pattern of the hemagglutinin gene of influenza B viruses isolated in Japan: cocirculating lineages in the same epidemic season. J Virol 64(6):2860-2865.

22. Knossow M, Gaudier M, Douglas A, Barrere B, Bizebard T, Barbey C, Gigant B, Skehel JJ. 2002. Mechanism of neutralization of influenza virus infectivity by antibodies. Virology 302(2):294-298.

23. Krystal M, Elliott RM, Benz EW, Jr., Young JF, Palese P. 1982. Evolution of influenza A and B viruses: conservation of structural features in the hemagglutinin genes. Proc Natl Acad Sci U S A 79(15):4800-4804.

24. Krystal M, Young JF, Palese P, Wilson IA, Skehel JJ, Wiley DC. 1983. Sequential mutations in hemagglutinins of influenza B virus isolates: definition of antigenic domains. Proc Natl Acad Sci U S A 80(14):4527-4531.

25. Lubeck MD, Schulman JL, Palese P. 1980. Antigenic variants of influenza viruses: marked differences in the frequencies of variants selected with different monoclonal antibodies. Virology 102(2):458-462.

26. Lugovtsev VY, Vodeiko GM, Levandowski RA. 2005. Mutational pattern of influenza B viruses adapted to high growth replication in embryonated eggs. Virus Res 109(2):149-157.

27. Lugovtsev VY, Vodeiko GM, Strupczewski CM, Ye Z, Levandowski RA. 2007. Generation of the influenza B viruses with improved growth phenotype by substitution of specific amino acids of hemagglutinin. Virology 365(2):315-323.

28. Macken C, Lu H, Goodman J, Boykin L. 2001. The value of a database in surveillance and Vaccine selection. In: Osterhaus A, Cox N, Hampson AW, editors. Options for the control of the influenza IV. Amsterdam: Elsevier Sciences. p 103-106.

29. Matsuzaki Y, Sugawara K, Takashita E, Muraki Y, Hongo S, Katsushima N, Mizuta K, Nishimura H. 2004. Genetic diversity of influenza B virus: the frequent reassortment and cocirculation of the genetically distinct reassortant viruses in a community. J Med Virol 74(1):132-140.

30. McCullers JA, Saito T, Iverson AR. 2004. Multiple genotypes of influenza B virus circulated between 1979 and 2003. J Virol 78(23):12817-12828.

31. McCullers JA, Wang GC, He S, Webster RG. 1999. Reassortment and

insertion-deletion are strategies for the evolution of influenza B viruses in nature. J Virol 73(9):7343-7348.

32. Muyanga J, Matsuzaki Y, Sugawara K, Kimura K, Mizuta K, Ndumba I, Muraki Y, Tsuchiya E, Hongo S, Kasolo FC, Numazaki Y, Nakamura K. 2001. Antigenic and genetic analyses of influenza B viruses isolated in Lusaka, Zambia in 1999. Arch Virol 146(9):1667-1679.

33. Nakagawa N, Kubota R, Maeda A, Nakagawa T, Okuno Y. 2000. Heterogeneity of influenza B virus strains in one epidemic season differentiated by monoclonal antibodies and nucleotide sequences. J Clin Microbiol 38(9):3467-3469.

34. Nakagawa N, Kubota R, Maeda A, Okuno Y. 2004. Influenza B virus victoria group with a new glycosylation site was epidemic in Japan in the 2002-2003 season. J Clin Microbiol 42(7):3295-3297.

35. Nakagawa N, Kubota R, Morikawa S, Nakagawa T, Baba K, Okuno Y. 2001a. Characterization of new epidemic strains of influenza B virus by using neutralizing monoclonal antibodies. J Med Virol 65(4):745-750.

36. Nakagawa N, Kubota R, Nakagawa T, Okuno Y. 2001b. Antigenic variants with amino acid deletions clarify a neutralizing epitope specific for influenza B virus Victoria group strains. J Gen Virol 82(Pt 9):2169-2172.

37. Nakagawa N, Kubota R, Nakagawa T, Okuno Y. 2003. Neutralizing epitopes specific for influenza B virus Yamagata group strains are in the 'loop'. J Gen Virol 84(Pt 4):769-773.

38. Nakagawa N, Kubota R, Okuno Y. 2005. Variation of the conserved neutralizing epitope in influenza B virus victoria group isolates in Japan. J Clin Microbiol

43(8):4212-4214.

39. Nakagawa N, Suzuoki J, Kubota R, Kobatake S, Okuno Y. 2006. Discovery of the neutralizing epitope common to influenza B virus victoria group isolates in Japan. J Clin Microbiol 44(4):1564-1566.

40. Nerome R, Hiromoto Y, Sugita S, Tanabe N, Ishida M, Matsumoto M, Lindstrom SE, Takahashi T, Nerome K. 1998. Evolutionary characteristics of influenza B virus since its first isolation in 1940: dynamic circulation of deletion and insertion mechanism. Arch Virol 143(8):1569-1583.

41. Oxford JS, Newman R, Corcoran T, Bootman J, Major D, Yates P, Robertson J, Schild GC. 1991. Direct isolation in eggs of influenza A (H1N1) and B viruses with haemagglutinins of different antigenic and amino acid composition. J Gen Virol 72 (Pt 1):185-189.

42. Oxford JS, Schild GC, Corcoran T, Newman R, Major D, Robertson J, Bootman J, Higgins P, al-Nakib W, Tyrrell DA. 1990. A host-cell-selected variant of influenza B virus with a single nucleotide substitution in HA affecting a potential glycosylation site was attenuated in virulence for volunteers. Arch Virol 110(1-2):37-46.

43. Pechirra P, Nunes B, Coelho A, Ribeiro C, Goncalves P, Pedro S, Castro LC, Rebelo-de-Andrade H. 2005. Molecular characterization of the HA gene of influenza type B viruses. J Med Virol 77(4):541-549.

44. Robertson JS, Bootman JS, Nicolson C, Major D, Robertson EW, Wood JM. 1990. The hemagglutinin of influenza B virus present in clinical material is a single species identical to that of mammalian cell-grown virus. Virology 179(1):35-40.

45. Robertson JS, Naeve CW, Webster RG, Bootman JS, Newman R, Schild GC. 1985.

Alterations in the hemagglutinin associated with adaptation of influenza B virus to growth in eggs. Virology 143(1):166-174.

46. Rota PA, Hemphill ML, Whistler T, Regnery HL, Kendal AP. 1992. Antigenic and genetic characterization of the haemagglutinins of recent cocirculating strains of influenza B virus. J Gen Virol 73 (Pt 10):2737-2742.

47. Rota PA, Wallis TR, Harmon MW, Rota JS, Kendal AP, Nerome K. 1990. Cocirculation of two distinct evolutionary lineages of influenza type B virus since 1983. Virology 175(1):59-68.

48. Saito T, Nakaya Y, Suzuki T, Ito R, Saito H, Takao S, Sahara K, Odagiri T, Murata T, Usui T, Suzuki Y, Tashiro M. 2004. Antigenic alteration of influenza B virus associated with loss of a glycosylation site due to host-cell adaptation. J Med Virol 74(2):336-343.

49. Schild GC, Oxford JS, de Jong JC, Webster RG. 1983. Evidence for host-cell selection of influenza virus antigenic variants. Nature 303(5919):706-709.

50. Schulze IT. 1997. Effects of glycosylation on the properties and functions of influenza virus hemagglutinin. J Infect Dis 176 Suppl 1:S24-28.

51. Shaw MW, Xu X, Li Y, Normand S, Ueki RT, Kunimoto GY, Hall H, Klimov A, Cox NJ, Subbarao K. 2002. Reappearance and global spread of variants of influenza B/Victoria/2/87 lineage viruses in the 2000-2001 and 2001-2002 seasons. Virology 303(1):1-8.

52. Skehel JJ, Stevens DJ, Daniels RS, Douglas AR, Knossow M, Wilson IA, Wiley DC. 1984. A carbohydrate side chain on hemagglutinins of Hong Kong influenza viruses inhibits recognition by a monoclonal antibody. Proc Natl Acad Sci U S A 81(6):1779-1783.

53. Skehel JJ, Wiley DC. 2000. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu Rev Biochem 69:531-569.

54. Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673-4680.

55. Verhoeyen M, Van Rompuy L, Jou WM, Huylebroeck D, Fiers W. 1983. Complete nucleotide sequence of the influenza B/Singapore/222/79 virus hemagglutinin gene and comparison with the B/Lee/40 hemagglutinin. Nucleic Acids Res 11(14):4703-4712.

56. Wang Q, Cheng F, Lu M, Tian X, Ma J. 2008. Crystal Structure of Unliganded Influenza B Virus Hemagglutinin. J Virol 82:3011-3020.

57. Webster RG, Berton MT. 1981. Analysis of antigenic drift in the haemagglutinin molecule of influenza B virus with monoclonal antibodies. J Gen Virol 54(Pt 2):243-251.

58. Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13(5):555-556.

59. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24(8):1586-1591.

60. Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155(1):431-449.

61. Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22(4):1107-1118.

Chapter Two

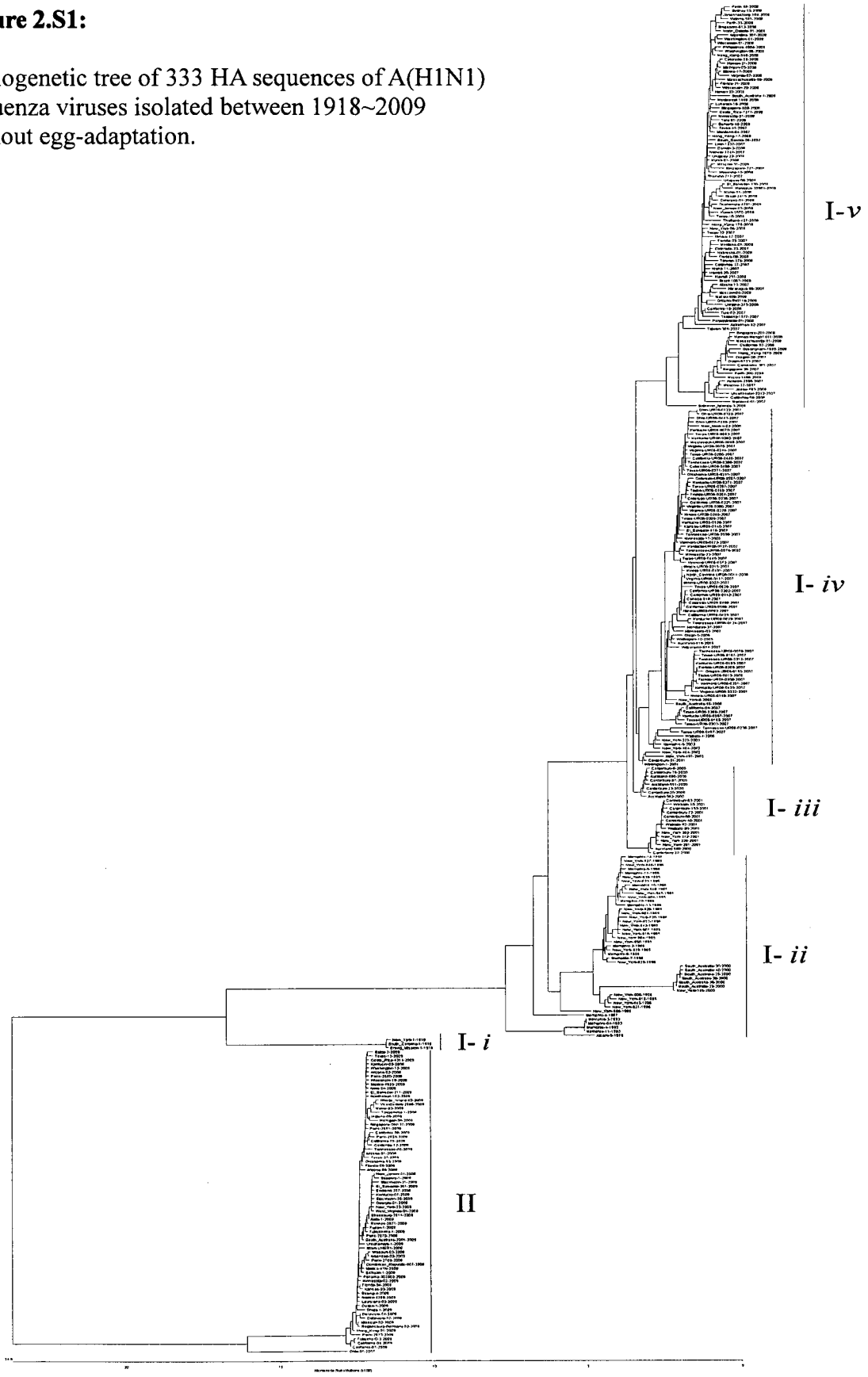
Evolutionary Trends of A(H1N1) Influenza Virus Hemagglutinin Since 1918

2.1 Abstract

The Pandemic (H1N1) 2009 is spreading to numerous countries and causing many human deaths. Although the symptoms in humans are mild at present, fears are that further mutations in the virus could lead to a potentially more dangerous outbreak in subsequent months. As the primary immunity-eliciting antigen, hemagglutinin (HA) is the major agent for host-driven antigenic drift in A(H3N2) virus. However, whether and how the evolution of HA is influenced by existing immunity is poorly understood for A(H1N1). Here, by analyzing hundreds of A(H1N1) HA sequences since 1918, we show the first evidence that host selections are indeed present in A(H1N1) HAs. Among a subgroup of human A(H1N1) HAs between 1918~2008, we found strong diversifying (positive) selection at HA_1 156 and 190. We also analyzed the evolutionary trends at HA_1 190 and 225 that are critical determinants for receptor-binding specificity of A(H1N1) HA. Different A(H1N1) viruses appeared to favor one of these two sites in host-driven antigenic drift: epidemic A(H1N1) HAs favor HA_1 190 while the 1918 pandemic and swine HAs favor HA_1 225. Thus, our results highlight the urgency to understand the interplay between antigenic drift and receptor binding in HA evolution, and provide molecular signatures for monitoring future antigenically drifted 2009 pandemic and seasonal A(H1N1) influenza viruses.

Figure 2.S1:

Phylogenetic tree of 333 HA sequences of A(H1N1) influenza viruses isolated between 1918~2009 without egg-adaptation.



2.2 Introduction and background

Since April 2009, a global outbreak caused by the swine-origin 2009 A(H1N1) influenza virus has spread to numerous countries [1,2,3,4,5,6,7,8,9,10], which warranted the declaration of "Pandemic (H1N1) 2009" by the World Health Organization on June 11, 2009. As of September 6, there had been over 277,607 infected individuals and at least 3,205 confirmed human deaths worldwide.

The Pandemic (H1N1) 2009 is not the first human pandemic caused by A(H1N1) influenza virus. During 1918~1919, the "Spanish" A(H1N1) influenza virus swept across the globe, infected ~25% of the entire population and claimed at least 50 million human lives worldwide [11]. In subsequent years, A(H1N1) influenza virus continued to circulate among humans and caused a number of severe outbreaks between 1920s and 1950s [12,13,14,15,16,17,18,19,20], in particular the A(H1N1) epidemic in 1950~1951 with mortality exceeding those of the 1957 "Asian" and 1968 "Hong Kong" pandemics [18,19,20]. In 1957, A(H1N1) influenza virus disappeared, replaced by a reassorted A(H2N2) influenza virus [21]. However, the A(H1N1) influenza virus reappeared in 1977, with a close genetic and antigenic similarity to those A(H1N1) viruses isolated in 1950 [22,23,24], and has co-circulated with A(H3N2) and type B influenza virus to cause seasonal human epidemics ever since.

The same 1918 pandemic A(H1N1) influenza virus was also spread to swine during 1918~1919, and became the so-called "classical" swine influenza [8,25,26,27], first isolated in North American in 1930 [26] and in Europe in 1976 [28,29]. In 1979, a novel lineage of avian-like A(H1N1) influenza virus, believed to have derived from closely related Eurasia avian influenza viruses, emerged in swine in Europe [30] and replaced the

classical swine A(H1N1) virus in this region [31,32,33]. These two classes of swine A(H1N1) viruses displayed different evolutionary trajectories [34]. In1998, a new triple-reassortant A(H3N2) virus, derived from North American avian, classical swine A(H1N1) and human A(H3N2) viruses, caused outbreaks in North American swine [35,36]. Mixing of the triple-reassortant H3N2 with established swine lineages gave rise to H1N1 and H1N2 reassortant swine viruses [37,38]. Since 2007, human infection caused by A(H1N1) swine virus has become a health concern in the United States [7].

The 2009 A(H1N1) influenza virus has its origin as a reassortant from a Eurasian avian-like swine A(H1N1) virus and a triple-reassortant virus circulating in North American swine [1,2,3,4,5,6,7,8]. As such, the 2009 A(H1N1) virus contains NA and M from Eurasian avian-like swine A(H1N1) virus, and the remaining genes from the triple-reassortant virus - PB2 and PA (avian virus), PB1 (human A(H3N2)), and HA, NP and NS (classical swine A(H1N1)) [1,2,3,4,5,6,7,8]. In a sense, we are continuingly living in a pandemic that started in 1918 [27]. Thus, it is not surprising for the similarly mild first waves of the 1918 and 2009 pandemics. Notably, the second wave of the "Spanish" influenza in the fall of 1918 became much more lethal, peaked within one month of the initial introductions in many communities [11]. This makes influenza virologists and healthcare officials fear that further mutations in the 2009 A(H1N1) virus could also lead to a potentially more dangerous second wave in subsequent months. Thus, in-depth studies on the 1918 pandemic strains as well as their post-pandemic decedents should provide critical new insights into the evolution of A(H1N1) in general, and the pandemic potential of the 2009 A(H1N1) in particular.

HA is one of the two major glycoproteins on the surface of influenza virus. It is

the primary antigen that elicits host immune response, and is also responsible for binding to sialic-acid receptors and for mediating viral entry into host cells [39]. The hallmarks of highly pathogenic influenza viruses among human population include easy human-to-human transmission as a result of high affinity of HA for human-like α(2,6) receptors, and significant difference in sequence and antigenicity of HA with existing seasonal and vaccine strains [1,2,39]. It has been demonstrated on 1918 A(H1N1) HA that HA_1 D190 and D225 are key determinants for effective binding to human-like α(2,6) receptors and consequently high infectivity of the virus among human population [11,40,41,42]. A single mutation D225G reduced the binding affinity for α(2,6) receptors [41,42] and the infectivity of the virus [40], while a double variant D190E/D225G rendered the HA non-binding to α(2,6) receptors [41,42] and the virus non-infectious [40].

In A(H3N2) virus, HA is the major agent for host-driven antigenic drift [43,44]. However, it is unclear whether or not and, if yes, how human immunity imposes selection on A(H1N1) HA. In order to address this critical issue, we undertook a systematic computational analysis of the evolution of H1 HA in the region of HA_1, which is the primary target for host immunity selection [43].

Recent years have witnessed an explosive expansion of available computational methods for phylogenetic analysis of selective pressure, including a variety of methods that look for different types of positive selection such as diversifying selection, toggling selection and directional selection [45,46,47,48,49,50,51,52,53,54] implemented in software packages such as HyPhy [55], MrBayes [56,57] and PAML [45]. Here we used PAML 4.0 [45] for calculation of heterogeneous selection pressure at each codon and

HyPhy [55] for directional selection in 335 non-egg-adapted and 32 egg-adapted human A(H1N1) HA sequences. These sequences were from A(H1N1) viruses isolated all around the globe between 1918~2009. In addition, we also analyzed 42 classical swine A(H1N1) HA sequences for their close relationship to the 2009 A(H1N1) HA.

In PAML 4.0 [45], a number of models are available: the branch models allow the ω ratio to vary among branches in the phylogenetic tree and can be used to detect positive selection on particular branches [46,58]; the site models allow the ω ratio to vary among sites and can be used to detect positive selection at particular sites [59,60]; the branch-site models allow the ω ratio to vary both among sites and among branches [61] and can be used to detect positive selection that affects only a few sites in a few branches.

In this analysis, a large dataset composed of over 300 sequences was used to ensure sufficient representative sequences for the total time span of 91 years, which made it impractical for the use of branch-site models in our calculations. However, by separating the sequences into distinct subgroups based on their phylogenetic relationship and applying the site models in PAML 4.0 [45], we successfully detected the branch and the specific sites therein that were under host-driven positive selection. Our study revealed differential evolutionary trends of A(H1N1) HA since 1918, which provided molecular signatures for monitoring future antigenically drifted 2009 pandemic and seasonal A(H1N1) influenza viruses.

2.3 Results and Discussion

A) Phylogenetic analysis of human A(H1N1) HA sequences since 1918

It is known that egg-adapted influenza viruses tend to have non-natural host-associated modifications at certain sites of HA sequences [62,63,64]. To eliminate the effects of such modifications in our analysis, we selected only 333 HA sequences of A(H1N1) viruses between 1918~2009 (as of July 10, 2009) with a well-documented record that they had never been passaged in chicken eggs at any stage. Furthermore, intragenic recombination may give rise to false positives in subsequent detection of positively selected codons [65], thus the Recombination Detection Program (RDP3) [66] was used to make sure that all HA sequences used in this study were free of recombination, agreeing with previous observations that intragenic recombination is rare for HA [67]. The nucleotide sequences of 333 A(H1N1) HAs in the region of HA_1 including the signal peptide, were analyzed by the ClustalW method [68]. The phylogeny tree suggested that these HA sequences belong to two major groups: the majority of HA sequences from 1918 to 2008 formed group I, and those of the 2009 A(H1N1) together with a strain isolated in 2007 formed group II (Fig. 2.S1). The separation of the 2009 A(H1N1) HAs from HAs of established human A(H1N1) viruses between 1918~2008, including the 1918 pandemic and the seasonal A(H1N1) viruses, was consistent with the proposed swine origin of HAs in these viruses [1,2,3,4,5,6,7,8]. The low sequence identity (~73%) between the 2009 A(H1N1) HA with seasonal and vaccine A(H1N1) HAs might explain why people were in general immunologically naïve to the former [8,69]. In fact, there did not exist cross-reactivity between the 2009 and seasonal A(H1N1) viruses [8], nor did the vaccination with recent (2005~2009) annual

vaccines provide immune protection against the 2009 A(H1N1) virus [69].

Table 2.1: The values of log-likelihood (l), d_N/d_S, and parameter estimates in CODEML analysis of human A(H1N1) HAs

Model	l	d_N/d_S	Parameters estimates
		I-*i* 1918~1919 (5 strains)[1]	
M0 (one-ratio)	-806.78	0.516	ω=0.516
M1a (nearly neutral)	-805.91	0.323	p_0=0.677 (p_1=0.323), ω_0=0 (ω_1=1)
M2a (positive selection)	-804.19	0.564	p_0=0.963, p_1=0 (p_2=0.037), ω_0=0 (ω_1=1), ω_2=15.421
M7 (beta)	-805.92	0.300	p=0.005, q=0.012
M8a (beta&ω=1)	-805.91	0.323	p_0=0.846 (p_1=0.154), p=0.005, q=0.020, ω_s=1
M8 (beta&ω>1)	-804.19	0.561	p_0=0.963 (p_1=0.037), p=0.005, q=7.228, ω_s=15.242
		I-*ii* 1979~2000 (45 strains)[2]	
M0 (one-ratio)	-2489.23	0.223	ω=0.223
M1a (nearly neutral)	-2486.25	0.242	p_0=0.855 (p_1=0.145), ω_0=0.113 (ω_1=1)
M2a (positive selection)	-2486.25	0.242	p_0=0.855, p_1=0.056 (p_2=0.089), ω_0=0.113 (ω_1=1), ω_2=1
M7 (beta)	-2485.63	0.232	p=0.327, q=1.074
M8a (beta&ω=1)	-2485.63	0.232	p_0=1 (p_1=0), p=0.327, q=1.074, ω_s=1
M8 (beta&ω>1)	-2485.63	0.232	p_0=1 (p_1=0), p=0.327, q=1.074, ω_s=1
		I-*iii* 2000~2001 (22 strains)[2]	
M0 (one-ratio)	-1686.21	0.279	ω=0.279
M1a (nearly neutral)	-1684.58	0.261	p_0=0.793 (p_1=0.207), ω_0=0.068 (ω_1=1)
M2a (positive selection)	-1683.09	0.287	p_0=0.994, p_1=0 (p_2=0.006), ω_0=0.225 (ω_1=1), ω_2=10.636
M7 (beta)	-1684.52	0.265	p=0.046, q=0.127
M8a (beta&ω=1)	-1684.58	0.261	p_0=0.793 (p_1=0.207), p=7.211, q=98.93, ω_s=1
M8 (beta&ω>1)	-1683.10	0.287	p_0=0.994 (p_1=0.006), p=28.774, q=99, ω_s=10.638
		I-*iv* 2001~2007 (89 strains)[2]	
M0 (one-ratio)	-2720.83	0.187	ω=0.187
M1a (nearly neutral)	-2709.20	0.181	p_0= 0.883 (p_1=0.117), ω_0=0.072 (ω_1=1)
M2a (positive selection)	-2708.27	0.187	p_0= 0.906, p_1=0.076 (p_2=0.018), ω_0=0.085 (ω_1=1), ω_2=1.915
M7 (beta)	-2708.99	0.183	p=0.137, q=0.612
M8a (beta&ω=1)	-2709.40	0.180	p_0=0.885 (p_1=0.115), p=7.871, q=98.995, ω_s=1
M8 (beta&ω>1)	-2708.82	0.186	p_0=0.973 (p_1=0.027), p=0.375, q=2.298, ω_s=1.936
		I-*v* 2006~2008 (100 strains)[2]	
M0 (one-ratio)	-3871.33	0.303	ω=0.303
M1a (nearly neutral)	-3813.28	0.241	p_0=0.828 (p_1=0.172), ω_0=0.082 (ω_1=1)
M2a (positive selection)	-3782.07	0.313	p_0=0.807, p_1=0.188 (p_2=0.005), ω_0=0.083 (ω_1=1), ω_2=11.142
M7 (beta)	-3814.21	0.246	p=0.139, q=0.427
M8a (beta&ω=1)	-3812.40	0.231	p_0=0.862 (p_1=0.138), p=0.497, q=3.969, ω_s=1
M8 (beta&ω>1)	-3781.55	0.305	p_0=0.994 (p_1=0.006), p=0.180, q=0.546, ω_s=10.554
		II 2007~2009 (74 strains)[2]	
M0 (one-ratio)	-2374.98	0.277	ω=0.277
M1a (nearly neutral)	-2373.20	0.282	p_0=0.922 (p_1=0.078), ω_0=0.222 (ω_1=1)
M2a (positive selection)	-2372.81	0.282	p_0=0.922, p_1=0.035 (p_2=0.042), ω_0=0.222 (ω_1=1), ω_2=1
M7 (beta)	-2373.21	0.281	p= 1.218, q=3.079
M8a (beta&ω=1)	-2374.03	0.282	p_0=0.945 (p_1=0.055), p=3.243, q=10.201, ω_s=1
M8 (beta&ω>1)	-2372.81	0.282	p_0=0.946 (p_1=0.054), p=3.158, q=9.900, ω_s=1

[1] Due to the inclusion of two partial sequences of A/London/1/1918 and A/London/1/1919 in this subgroup, the analysis was performed on a total of 187 amino-acid residues that covered the antigenic and receptor-binding sites in the region of HA_1 (51~237) [11]. [2] The analysis was performed on the first 340 residues of HA_1 including the signal peptide.

Table 2.2: LRT tests for HA_1 sequences of human A(H1N1) influenza viruses

	LRT (M2a – M1a) ($2\Delta l$) (p-values)[1]	LRT (M8 – M7) ($2\Delta l$) (p-values)[1]	LRT (M8 – M8a) ($2\Delta l$) (p-values)[2]
I-*i* 1918~1919 (5 strains)	3.44 (0.1791)	3.46 (0.1773)	3.44 (0.0318)
I-*ii* 1979~2000 (45 strains)	0	0	0
I-*iii* 2000~2001 (22 strains)	2.98 (0.2254)	2.84 (0.2417)	2.96 (0.0427)
I-*iv* 2001~2007 (89 strains)	1.86 (0.3946)	0.34 (0.8437)	1.16 (0.1407)
I-*v* 2006~2008 (100 strains)	62.42 (0.0000)	65.32 (0.0000)	61.70 (0.0000)
II 2007~2009 (74 strains)	0.78 (0.6771)	0.80 (0.6703)	2.44 (0.0591)

[1] We used the degree of freedom of 2 for these LRT tests that is expected to be too conservative.

[2] The p-values were calculated from χ^2 distribution using degree of freedom of 1 that was then divided by a factor of 2 for the mixture distribution, as suggested by the author of PAML 4.0.

B) Evidence for host-driven antigenic drift in human A(H1N1) HAs

In order to understand whether host-driven antigenic drift is imposed on the evolution of HA_1 of A(H1N1) virus, we used likelihood ratio tests (LRT) in the software package PAML 4.0 [45] to identify the presence or absence of *positive selection.* In this context, positive selection referred to a significant excess of amino-acid altering (non-synonymous) substitutions over silent (synonymous) substitutions in nucleotide sequences. Large LRT values (or small p-values) between alternative models and null models, such as M2a vs. M1a, M8 vs. M7, or M8 vs. M8a, led to the rejection of the null

models.

Since HA sequences of group I was further divided into five subgroups (Fig. 2.S1), the PAML calculation was carried out on each of these five subgroups and on group II (Table 2.1). For group I-*i* that included three 1918 pandemic A(H1N1) HAs, in order to increase the sample size, we also included two partial sequences, A/London/1/1918 and A/London/1/1919 [11]. Except for the subgroup I-*v*, all other subgroups of group I had very low LRT values and large *p*-values (Table 2.1, 2.2), indicating predominantly neutral or purifying selection. These results were consistent with the overall low prevalence of A(H1N1) virus during the period of 1979~2006 [70], and agreed well with a previous study that focused on 1995~2005 A(H1N1) isolates where no positive selection was detected [71]. In sharp contrast, group I-*v* 2006~2008 had $\omega>10$ and LRT>60, which provided strong evidence for positive selection (Table 2.1, 2.2) and agreed with the necessity to update the A(H1N1) vaccine strain using A/Brisbane/59/07 for the 2008~2009 season. Group II including 73 HAs of 2009 A(H1N1) and one of 2007 A(H1N1) also had a very low LRT rate ratio (Table 2.1, 2.2). Given the largely nonexistence of human immunity against the 2009 A(H1N1), the lack of positive selection among group II was expected. However, with more mild infections rapidly propagating among human population in the first wave, the gradually established human immunity might drive positive selection in future isolates of 2009 A(H1N1) strains.

Table 2.3: Codons under positive selection in HA_1 of human A(H1N1) influenza viruses

		Positively selected sites[1]
I-*v* 2006~2008 (100 strains)	M2a	156 (97.5%), 190 (100%)
	M8	156 (99.7%), 190 (100%)

[1]Positively selected sites from PAML 4.0 [45] using Bayes Empirical Bayes analysis [72]. Only codons with greater than 95% posterior probabilities to be under positive selection were listed with the corresponding posterior probabilities shown in parentheses.

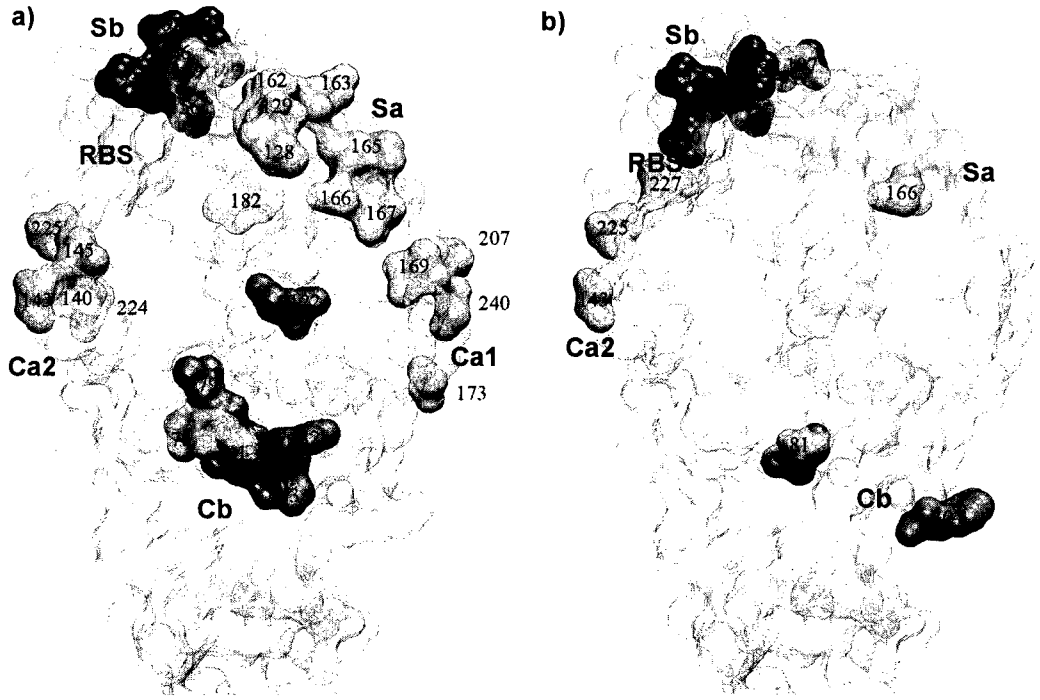


Figure 2.1: Antigenic structure and positive selection of A(H1N1) HA.

a) Antigenic structure of A/PR/8/34 (H1N1) HA (PDB accession code 1RU7 [78]). Five antigenic sites were identified by using a large number of monoclonal antibodies [73,74]: Sa (cyan), Sb (red), Ca1 (yellow), Ca2 (green), Cb (blue), using H3 HA numbering. The receptor-binding site (RBS) was labeled for reference. **b)** Codons on A(H1N1) HA that were identified to be under various selection in PAML and HyPhy analysis.

C) Identification of positively selected codons in human A(H1N1) HAs

In order to understand how H1 HA sequences were positively selected by human existing immunity, the CODEML [72] program in PAML 4.0 was used on subgroup I-*v* in which about 0.6% codons were found to be under positive selection (Table 2.1, 2.2). Both M2a and M8 models identified HA_1 156 and 190 with greater than 95% posterior probabilities to be under positive selection (Table 2.3). In previous studies, the antigenic structure of H1 HA (A/PuertoRico/8/1934) had been determined to include five distinct antigenic sites on the globular domain: Sa, Sb, Ca1, Ca2 and Cb [73,74] (Fig. 2.1a). Both of these positively selected codons were located on the site Sb (Fig. 2.1b). The focus of positive selection on the Sb antigenic site was consistent with a cross-reactivity analysis of various epidemic H1N1 strains using monoclonal antibodies that it was under much higher pressure for mutations [74].

HA_1 138, 186, 190, 194, 225, 226 and 228 had been previously shown to affect receptor binding to H1 HA [75,76]. Among them, two residues, HA_1 190 and 225, play predominant roles in determining the receptor-binding specificity of H1 HA: D190/D225 for α(2,6) receptors in humans, D190/G225 for α(2,6) and α(2,3) receptors in swine, and E190/G225 for α(2,3) receptors in avian [11,39,40,41,42,76]. Although changes at these two sites had been previously reported to cause antigenic drift in A(H1N1) epidemic strains [77], it was a somewhat common belief that key determinants of receptor-binding specificity are in general not subject to selection. Thus, the strong positive selection at HA_1 190 within subgroup I-*v* is quite unexpected.

D) Positive selection of egg-adapted human A(H1N1) HAs during 1933~1979

Table 2.4: The values of log-likelihood (l), d_N/d_S, and parameter estimates in CODEML analysis of egg-adapted human A(H1N1) HAs between 1933-1979

Model	l	d_N/d_S	Parameters estimates
1933~1979 (32 strains)[1]			
M0 (one-ratio)	-3336.01	0.411	ω=0.411
M1a (nearly neutral)	-3283.75	0.336	p_0=0.705 (p_1=0.295), ω_0=0.057 (ω_1=1)
M2a (positive selection)	-3275.24	0.454	p_0=0.721, p_1=0.234 (p_2=0.046), ω_0=0.079 (ω_1=1), ω_2=3.571
M7 (beta)	-3285.56	0.345	p=0.068, q=0.129
M8a (beta&ω=1)	-3283.78	0.336	p_0=0.705 (p_1=0.295), p=6.100, q=99, ω_s=1
M8 (beta&ω>1)	-3275.41	0.452	p_0=0.942 (p_1=0.058), p=0.206, q=0.534, ω_s=3.283
1947~1957 (12 strains)[1]			
M0 (one-ratio)	-2068.01	0.435	ω=0.435
M1a (nearly neutral)	-2056.27	0.337	p_0=0.689 (p_1=0.311), ω_0=0.038 (ω_1=1)
M2a (positive selection)	-2046.27	0.501	p_0=0.967, p_1=0 (p_2=0.033), ω_0=0.256 (ω_1=1), ω_2=7.651
M7 (beta)	-2056.44	0.324	p=0.012, q=0.023
M8a (beta&ω=1)	-2056.28	0.337	p_0=0.688 (p_1=0.312), p=3.886, q=99, ω_s=1
M8 (beta&ω>1)	-2046.29	0.501	p_0=0.967 (p_1=0.033), p=34.141, q=99, ω_s=7.667
1948~1979 (17 strains)[1]			
M0 (one-ratio)	-1759.58	0.385	ω=0.385
M1a (nearly neutral)	-1751.46	0.260	p_0=0.740 (p_1=0.260), ω_0=0 (ω_1=1)
M2a (positive selection)	-1747.17	0.408	p_0=0.794, p_1=0.168 (p_2=0.039), ω_0=0 (ω_1=1), ω_2=6.226
M7 (beta)	-1751.61	0.300	p=0.005, q=0.012
M8a (beta&ω=1)	-1751.46	0.260	p_0=0.740 (p_1=0.260), p=0.005, q=2.350, ω_s=1
M8 (beta&ω>1)	-1747.19	0.411	p_0=0.969 (p_1=0.031), p=0.006, q=0.025, ω_s=6.964

[1] The analysis was performed on the first 337 residues of HA_1 including the signal peptide.

Table 2.5: LRT tests and codons under positive selection for HA_1 sequences of egg-adapted human A(H1N1) influenza viruses between 1933-1979

			LRT ($2\Delta l$) (p-values)	Positively selected sites[1]
1933~1979 (32 strains)	M2a	M2a-M1a	17.02 (0.0002)	**77 (95.8%),** 225 (98.8%)
	M8	M8-M7	20.30 (0.0000)	**77 (98.7%),** 225 (99.6%), **227 (97.7%)**
		M8-M8a	16.74 (0.0000)	
1947~1957 (12 strains)	M2a	M2a-M1a	20.0 (0.0000)	**143 (99.3%), 264 (99.6%)**
	M8	M8-M7	20.30 (0.0000)	**143 (99.6%), 166 (95.1%), 264 (99.7%)**
		M8-M8a	19.98 (0.0000)	
1948~1979 (17 strains)	M2a	M2a-M1a	8.58 (0.0137)	225 (99.1%)
	M8	M8-M7	8.84 (0.0120)	225 (99.8%)
		M8-M8a	8.54 (0.0017)	

[1]Positively selected sites from PAML 4.0 [45] using Bayes Empirical Bayes analysis [72]. Only codons with greater than 95% posterior probabilities to be under positive selection were listed with the corresponding posterior probabilities shown in parentheses. Highlighted in bold were codons that were not associated with egg-adapted substitutions [62,63,64].

To compensate for the lack of non-egg-adapted human A(H1N1) HAs for the period of 1933~1978, we separately collected a total of 32 different egg-adapted A(H1N1) HA sequences between 1933~1979 that were free of sequence ambiguity (Fig. 2.S2). These sequences as a group were analyzed by PAML 4.0, as well as two subgroups that covered the periods of 1947~1957 (12 sequences) and 1948~1979 (17 sequences) (Table 2.4, 2.5), keeping in mind of the egg-adapted mutations at HA_1 138, 144, 163, 189, 190, 225, and 226 [62,63,64]. The two subgroups 1947~1957 and 1948~1979 represented A(H1N1) viruses circulating in the 1950s and in the 1970s upon its reemergence in 1977, respectively. Given the close genetic and antigenic similarity of the reappeared A(H1N1) influenza virus in 1977 with the A(H1N1) viruses isolated in 1950 [22,23,24], it was of particular interest to investigate whether different evolutionary trends were adopted by the 1947~1957 and 1948~1979 subgroups.

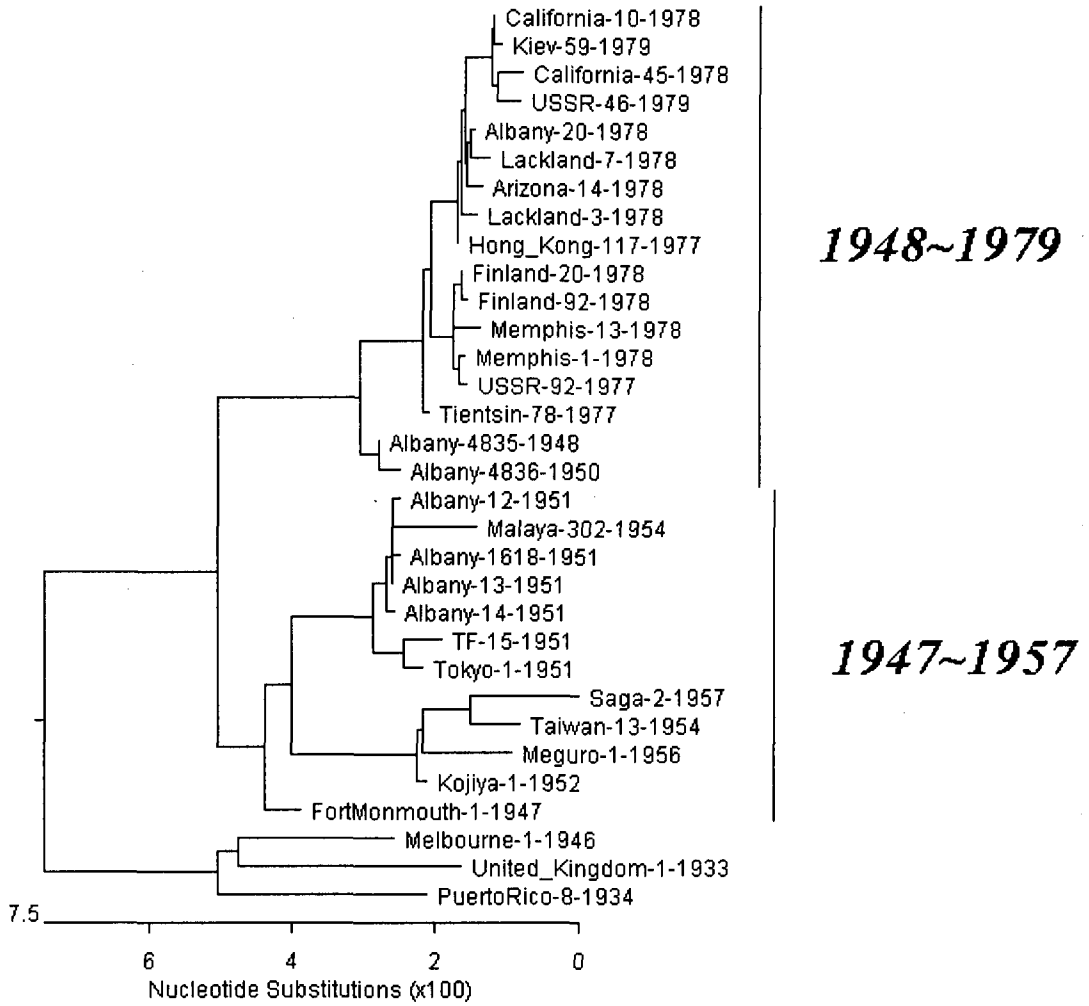


Figure 2.S2: Phylogenetic tree of 32 HA sequences of egg-adapted human A(H1N1) influenza viruses isolated between 1933~1979.

For both the entire group 1933~1979 and the subgroup 1947~1957, comparisons of M2a-M1a, M8-M7, or M8-M8a yielded large LRT values and very small *p*-values, suggesting the presence of positive selection at about 5% and 3% codons, respectively (Table 2.4, 2.5). However, it is noteworthy that the subgroup 1948~1979 had much smaller LRT values, suggesting that the positive pressure of the entire group 1933~1979 be mostly from the contribution of the subgroup 1947~1957.

We further employed the CODEML in PAML 4.0 to analyze the positively

selected codons in each group. The results were shown in Table 2.5 where highlighted in bold were the codons not known to be possible egg-adapted mutations (HA_1 138, 144, 163, 189, 190, 225, and 226) [62,63,64]. For the entire group 1933~1979, HA_1 77, 225 and 227 were found to be under positive selection with greater than 95% posterior probability in model M8 (Fig. 2.1b). They were located in the antigenic sites Cb (HA_1 77) and Ca2 (HA_1 225 and 227), respectively (Fig. 2.1b). In addition, for the subgroup 1948~1979, HA_1 225 was found to be under positive selection with greater than 99% posterior probability in both models M2a and M8. However, given the fact that these HAs were from egg-adapted A(H1N1) viruses in which HA_1 225 was one of the most frequently changed site [62,63,64], and the predominant residue at this site (Table 2.6), G225, was commonly found in swine and avian A(H1N1) HAs, it was possible that the changes at HA_1 225 was due to positive selection imposed by adaptation in eggs. At posterior probability of 90%, HA_1 138 and 189 were positively selected as well, however, both sites were involved in egg-adapted substitutions [62,63,64]. In sharp contrast, however, HA_1 143, 166 and 264 in the subgroup 1947~1957 were found to be under positive selection (Table 2.5), none of which was among the previously identified egg-adapted mutations. Antigenically, these codons were located in the antigenic sites Ca2, Sa and Cb, respectively (Fig. 2.1b). For their relatively distant location from the receptor-binding site, HA_1 143, 166 and 264 are probably mutations driven by existing human immunity for antibody escape.

Table 2.6: Codons at HA_1 190 and 225 in human and swine A(H1N1) influenza viruses

	D190	Non-D190	D225	G225	Non-D225/G225
Human 1979~2008 Epidemic (575 sequences)	477 (83.0%)*	98 (17.0%)	565 (98.2%)	2 (0.4%)	8 (1.4%)
Human 1918 Pandemic (5 sequences)	5 (100%)	0	3 (60%)	2 (40%)	0
Human 2009 Pandemic (73 sequences)	73 (100%)	0	69 (94.5%)	1 (1.4%)	3 (4.1%)
Human 1947~1957 (12 strains) (egg-adapted)	9 (75%)	3 (25%)	2 (16.7%)	10 (83.3%)	0
Human 1948~1979 (17 strains) (egg-adapted)	16 (94.1%)	1 (5.9%)	4 (23.5%)	13 (76.5%)	0
Swine 1990~2009 (42 sequences)	41 (97.6%)	1 (2.4%)	28 (66.6%)	12 (28.6%)	2 (4.8%)

*The number of cases that a particular type of residues occurs at each site. Shown in parenthesis was the occurrence in percentage.

Thus, there appeared to have different evolutionary patterns for the subgroup 1947~1957 circulating in the 1950s and the subgroup 1948~1979 circulating mostly in the 1970s. The former subgroup was subjected to positive selection pressure at HA_1 143, 166 and 264 (Table 2.5), and had a much larger variability at HA_1 190, with 25% being non-D190 (Table 2.6). In marked contrast, the latter subgroup was probably not under host-driven positive selection in humans and had highly conserved HA_1 190 (94.1% being D190).

E) Evolution of swine A(H1N1) HAs during 1990~2009

Table 2.7: The values of log-likelihood (l), d_N/d_S, and parameter estimates in CODEML analysis of swine A(H1N1) HAs between 1990-2009

Model	l	d_N/d_S	Parameters estimates	LRT (2Δl) (*p*-values)
M0 (one-ratio)	-6021.08	0.158	ω=0.158	
M1a (nearly neutral)	-5949.30	0.217	p_0=0.864 (p_1=0.136), ω_0=0.094 (ω_1=1)	LRT (M2a-M1a) = 1.24 (0.5379)
M2a(positive selection)	-5948.68	0.224	p_0=0.864, p_1=0.134 (p_2=0.002), ω_0=0.095 (ω_1=1), ω_2=3.682	
M7 (beta)	-5925.79	0.174	p= 0.381, q=1.768	LRT (M8-M7) = 6.78 (0.0337)
M8a (beta&ω=1)	-5924.10	0.171	p_0=0.970 (p_1=0.030), p=0.460, q=2.617, ω_s=1	LRT (M8-M8a) = 3.40 (0.0326)
M8 (beta&ω)	-5922.40	0.177	p_0=0.995 (p_1=0.005), p=0.413, q=2.052, ω_s=2.546	

The analysis was performed on the first 338 residues of HA_1 including the signal peptide.

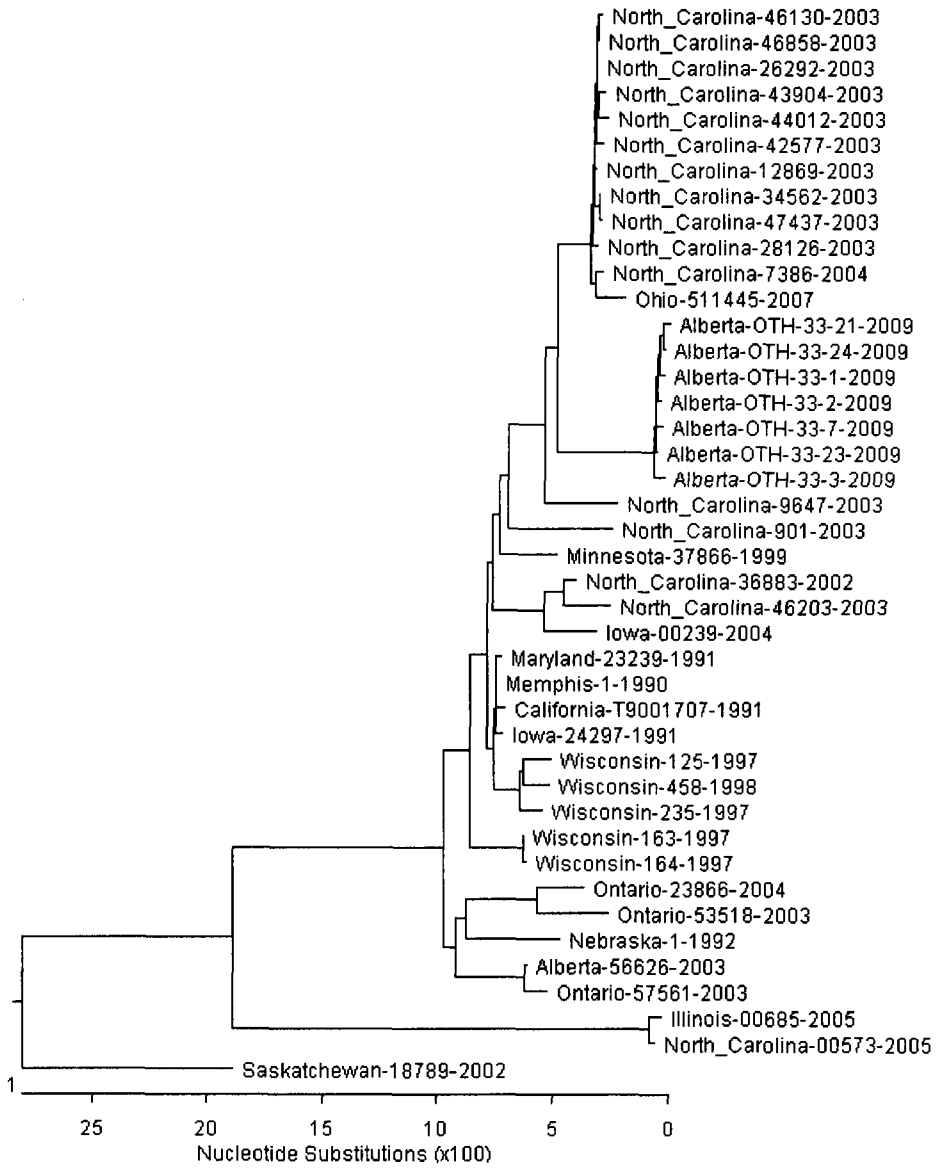


Figure 2.S3: Phylogenetic tree of 42 HA sequences of swine A(H1N1) influenza viruses isolated between 1990~2009.

Given the swine origin of the 2009 pandemic A(H1N1) HA, we also analyzed 42 non-redundant, non-ambiguous swine A(H1N1) HA sequences during 1990~2009 available from GISAID/Epifludb (Table 2.7, Fig. 2.S3). The reason that we focused on

this period was mainly for the antigenic stasis of swine A(H1N1) until 1998 [8] since the introduction of the 1918 "Spanish" A(H1N1) virus into swine [1,2,3,4,5,6,7,8]. Overall, the alternative models M2a and M8 fitted the data only marginally better than the null models M1a, M7 and M8a, respectively (Table 2.7). Thus, it seemed that swine A(H1N1) HAs during 1990~2009 were not subjected to strong host-driven positive selection.

Table 2.8: Directional selection analysis on human A(H1N1) HAs

	Tree L	Residue	*P*-Value	Bias	Proportion (%)	No. of Sites
Human I-*v* (100 sequences)	0.474	T	0.0002	32.995	5.4	2
		R	0.0002	12.583	11.8	1
		V	0.0004	34.478	4.0	2
		K	0.0023	29.301	7.2	1
Human 1933~1979 (32 strains) (egg-adapted)	0.613	D	0.0000	78.148	3.4	1
Human 1948~1979 (17 strains) (egg-adapted)	0.091	D	0.0007	133.611	5.1	1

Table 2.9: Sites found to be under directional selection in human A(H1N1) HAs

	Sites	Composition	Root	Preferred	Inferred Substitutions
Human I-*v* (100 sequences)	143	$V_{99}T_1$	V	V	$T\rightarrow_1 V$
	156	$G_{90}R_9E_1$	R	R	$G\rightarrow_1 E$, $G\rightarrow_7 R$
	158	$N_{96}K_4$	N	K	$N\rightarrow_4 K$
	190	$D_{67}N_{25}V_8$	D	V	$D\rightarrow_{16} N$, $D\rightarrow_4 V$
	193	$A_{52}T_{48}$	A	T	$A\rightarrow_6 T$
	197	$T_{82}K_{18}$	T	T	$K\rightarrow_2 T$
Human 1933~1979 (32 strains) (egg-adapted)	225	$G_{23}D_9$	D	D	$D\rightarrow_1 G$, $G\rightarrow_6 D$
Human 1948~1979 (17 strains) (egg-adapted)	225	$G_{13}D_4$	G	D	$G\rightarrow_4 D$

F) Directional evolution of human A(H1N1) HAs

In order to test whether directional evolution of protein sequences existed in the evolution of human A(H1N1) HAs, we employed a maximum likelihood method

developed by Kosakovsky Pond and colleagues [49]. In each subgroup, we used the oldest HA sequence as the root. In agreement with CODEML analysis reported in previous sections, among all non-egg adapted human A(H1N1) HAs, directional evolution was only identified in the subgroup I-*v*, at sites HA_1 143, 156, 158, 190, 193 and 197 (Table 2.8, 2.9). HA_1 143 belonged to the antigenic site Ca2 of A(H1N1) HA, whilst all other sites were located in the antigenic site Sb (Fig. 2.1b). Among these sites, HA_1 156, 190 and 193 were identified by CODEML in PAML 4.0 to be under positive selection with 99.7%, 100%, and 80.9% posterior probability in model M8, respectively (Table 2.3). In previous structural studies, residue HA_1 190 in 1934 human A(H1N1) HA and HA_1 190 and 193 in 1930 swine A(H1N1) HA were found to directly interact with bound human-like α(2,6)-receptors [78]. Thus, it remains to be investigated the impacts of directional evolution at HA_1 190 and 193 on receptor binding and antigenic drift.

We also performed directional evolution study on egg-adapted human A(H1N1) HA sequences, and found that in both the entire group 1933~1979 and the subgroup 1948~1979, multiple favored mutations of D225→G and G225→D were detected (Table 2.8, 2.9). Given its involvement in egg-adaptation, the directional evolution at HA_1 225 may be the consequence of egg-adaptation. In contrast, no residues in the subgroup 1947~1957 were identified to be under directional selection.

G) Evolution of human and swine A(H1N1) HAs at HA_1 190 and 225

For their predominant roles in determining receptor-binding specificity of A(H1N1) HA, and the positive selection on HA_1 190 in the subgroup I-*v*, we further investigated the evolution of HA_1 190 and 225 in A(H1N1) strains during 1918~2009. These included 653 non-egg-adapted HAs (five pandemic HAs from 1918~1919, 575 epidemic HAs from 1979~2008, and 73 pandemic HAs from 2009), and 42 swine HAs (Table 2.6). For the 575 epidemic HAs, HA_1 190 was highly variable (17.0% sequences did not have D190), while HA_1 225 was more conserved (only 1.8% sequences did not have D225) (Table 2.6). Among all the deviations (a total of 107 cases) from the ideal D190/D225 combination for human A(H1N1) viruses, two predominant ones were N190/D225 (69.2%) and V190/D225 (19.6%). At present, we don't know the exact effects of these mutations, or in combination with other concurring mutations at or around the receptor-binding site, on binding to human receptors. Further experiments are needed to clarify these issues. However, in previous studies, a single mutation D190N of A(H1N1) HA was shown to result in a lower binding affinity for human-like $\alpha(2,6)$ receptors, and a higher binding affinity for avian-like $\alpha(2,3)$ receptors [64].

The five HA sequences retrieved from victims of 1918 "Spanish" A(H1N1) influenza virus shared 98.9% to 99.8% sequence identity [11]. Among them, there were two non-synonymous substitutions of D225G, one in A/New York/1/1918 and the other one in A/London/1/1919 (Table 2.6). The HAs harboring the mutation D225G had reduced binding affinity for human receptors [11,40,41].

In the 73 HA sequences from the 2009 pandemic A(H1N1), D190 was strictly conserved, while D225 was 94.5% conserved (Table 2.6). At HA_1 225, the deviations were 1.1% for G225 and 3.3% for E225. Thus, the complete conservation at HA_1 190 and the nearly complete conservation at HA_1 225 were consistent to the importance of these residues in allowing for binding to human-like $\alpha(2,6)$ receptors [40,41], supporting the substantially higher human-to-human transmissibility of the 2009 A(H1N1) virus than seasonal A(H1N1) viruses [5,8].

Therefore, there were two distinct evolutionary trends in host-driven antigenic drift of human A(H1N1) HAs at residues in the receptor-binding site: the 1918 pandemic HAs underwent antigenic drift at HA_1 225, while the epidemic HAs undertook antigenic drift at HA_1 190. In the absence of selection, the 2009 A(H1N1) viruses were highly conserved at both HA_1 190 and 225, which was distinct from those two host-selected evolutionary trends (Table 2.6). With gradually established immunity among human population, we wondered how the 2009 A(H1N1) virus would undergo antigenic drift in the months to come. Thus, we also looked at the conservation at HA_1 190 and 225 in 42 swine A(H1N1) HA sequences (Table 2.6). Surprisingly, among these sequences, D190 was conserved at 97.6%, while D225 and G225 were observed at 66.6% and 28.6%, respectively. The similarly high variability of HA_1 225 in swine A(H1N1) HAs with that of 1918 pandemic HAs was consistent with the relative antigenic stasis of swine A(H1N1) until 1998 [8] and agreed well with the suggestion that the introduction of the 2009 pandemic A(H1N1) virus into humans be a single event or multiple events of similar viruses [1,2,3,4,5,6,7,8].

The deviations from the ideal D190/D225 combination in A(H1N1) HAs might result in reduced binding to human receptors [11,41,42,64,79]. However, two possibilities, which are not mutually exclusive, may explain the fact that mutations are frequently observed at these two sites: one is that other concurring mutations at or around the receptor-binding site may sufficiently maintain the receptor binding affinity so that the overall binding affinity is largely unaffected; the second is that the gain in evading antibody neutralization far overweighs the reduction in receptor binding. Due to the overlapping locations of the ever-changing antigenic sites and the more-conserved receptor-binding site of HA, there is a constant dilemma of whether or not a residue at the receptor-binding site should change. Although the involvement of residues in antigenic drift that are critical for receptor binding was also observed in HAs of other types and subtypes including influenza B virus HA [80], H3 [43,44]and H5 HA [44,81], the interplay between these two opposing forces in HA evolution is still very poorly understood. Although previous studies on A(H3N2) HAs suggested covariation of antigenicity and receptor-binding specificity as a possible mechanism for the antigenic differences observed in viruses propagated in different cells [82], questions such as how residues involved in receptor binding are actively utilized for antigenic drift in influenza evolution in the same hosts need to be urgently addressed in order for us to comprehend the powerful strategies that the virus employs for recurring influenza infections.

H) Implications for the 2009 pandemic

By analyzing hundreds of A(H1N1) HA sequences between 1918~2009, our study

revealed positive selection in the subgroup I-*v* of A(H1N1) HAs. The positively selected codons were located at HA_1 156 and 190 in the Sb antigenic site [83]. It was surprising that HA_1 190, which is critical for receptor-binding specificity of A(H1N1) HAs, was also under positive selection. Through further analysis of HA_1 190, together with HA_1 225, the other critical determinant for receptor-binding specificity of A(H1N1), we found that the epidemic HAs and the 1918 pandemic and swine HAs favored one of these two sites for antigenic drift. Whether the 2009 pandemic A(H1N1) HA will adopt any of these two trends, or use a novel mechanism that does not involve HA_1 190 and 225, will unfold in the coming months. If the latter is to be used, the 2009 A(H1N1) viruses may maintain their intrinsic high transmissibility, which, together with mutations in other genes such as NS1 and PB1-F2 with signatures of elevated pathogenicity [1,2], may suffice a new disastrous pandemic in the near future.

2.4 Materials and Methods

A) Phylogenetic analysis of A(H1N1) HAs

We obtained all available HA sequences (over 1,000) of non-egg-adapted A(H1N1) viruses for the period of 1918~2009 (as of July 10, 2009) from GISAID/Epifludb. We then removed the sequences with one or more ambiguous nucleotide sequences within the HA_1 region and deleted identical sequences. This gave us a dataset of 652 HA sequences that included three 1918 pandemic HAs, 575 epidemic HAs from 1979~2008 that collectively formed group I, and 73 pandemic HAs from 2009 and one HA from 2007 that belonged to group II. To facilitate the speed of computing, we further removed closely related sequences and obtained a dataset of 333 HA sequences. The program RDP3 (http://darwin.uvigo.es/rdp/rdp.html) [66] was used to make sure that no recombination was present in any of these HA sequences. The ClustalW method [68] with the MEGALIGN program of DNASTAR package (www.dnastar.com) was used for phylogenetic analysis of H1 HA sequences in the region of HA_1 (Fig. 2.S1).

Due to the historic use of eggs for amplification of influenza viruses before sequencing, there presented a vacuum in sequence for non-egg-adapted A(H1N1) viruses between 1919 and 1979. In order to gain insights into the evolution of A(H1N1) viruses for this period, we separately collected a total of 32 different egg-adapted A(H1N1) HA sequences between 1933~1979 that were free of sequence ambiguity (Fig. 2.S2). These sequences were similarly analyzed while keeping in mind of the possible egg-adapted mutations at HA_1 138, 144, 163, 189, 190, 225, and 226 [62,63,64].

In order to compare the evolution of swine A(H1N1) HA sequences, we also retrieved 42 unique swine H1 HA sequences for the period of 1990~2009 that were free of ambiguous nucleotide sequences (Fig. 2.S3). The reason that we focused on 1990~2009 was that previous studies suggested that swine A(H1N1) viruses be antigenically stable for the period of 1930 to 1990s [84].

B) Analysis of positive selection by PAML 4.0

The site-specific models implemented in the CODEML program in PAML 4.0 [45] was used to calculate heterogeneous selection pressure at amino-acid positions [45,54,72,85]. The models used in this study were M0, M1a, M2a, M7 and M8. M1a (nearly neutral), M7 (beta) and M8a (beta and ω=1) were null models that did not support $\omega > 1$. In contrast, the alternative models M2a (positive selection) and M8 (beta and ω), compared to M1a and M7 respectively, each had an additional class that allowed $\omega > 1$. Likelihood ratio tests (LRT) comparing M2a versus M1a, M8 versus M7, and M8 versus M8a provided test for the existence of positive selection. In the test, twice the log likelihood difference, $2\Delta l = 2(l_1-l_0)$, was calculated where l_1 and l_0 were the log likelihoods for the alternative model and null model, respectively. A larger value of LRT over those of χ^2 distribution led to rejection of the null models [72]. In order to calculate the codon-substitution models for heterogeneous selection pressure at each codon, the Bayes Empirical Bayes (BEB) analysis implemented in CODEML [72] was used, which has been shown to yield robust results even for small datasets. For all calculations, multiple runs, each with different initial parameter values, were performed to ensure optimization and convergence.

C) Directional evolution of protein sequences using HyPhy

Each group of A(H1N1) HA sequences aligned by the ClustalW method (Fig. 2.S1, 2.S2, 2.S3) was input to the PhyML program [86] to generate an unrooted phylogenetic tree, which was then rooted using the Treeview software [87] by selecting the oldest sequence in each group as the root/ancestor. This rooted phylogenetic tree was used for directional evolution of protein sequences [49] implemented in the HyPhy [55] software package.

2.5 References:

1. Wang TT, Palese P (2009) Unraveling the mystery of swine influenza virus. Cell 137: 983-985.
2. Neumann G, Noda T, Kawaoka Y (2009) Emergence and pandemic potential of swine-origin H1N1 influenza virus. Nature 459: 931-939.
3. Peiris JS, Poon LL, Guan Y (2009) Emergence of a novel swine-origin influenza A virus (S-OIV) H1N1 virus in humans. J Clin Virol 45: 169-173.
4. Dawood FS, Jain S, Finelli L, Shaw MW, Lindstrom S, et al. (2009) Emergence of a novel swine-origin influenza A (H1N1) virus in humans. N Engl J Med 360: 2605-2615.
5. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, et al. (2009) Pandemic Potential of a Strain of Influenza A (H1N1) : Early Findings. Science 324: 1557-1561.
6. Solovyov A, Palacios G, Briese T, Lipkin WI, Rabadan R (2009) Cluster analysis of the origins of the new influenza A(H1N1) virus. Euro Surveill 14.
7. Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, et al. (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459: 1122-1125.
8. Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, et al. (2009) Antigenic and Genetic Characteristics of Swine-Origin 2009 A(H1N1) Influenza Viruses Circulating in Humans. Science.
9. Munster VJ, de Wit E, van den Brand JM, Herfst S, Schrauwen EJ, et al. (2009) Pathogenesis and Transmission of Swine-Origin 2009 A(H1N1) Influenza Virus in

Ferrets. Science.

10. Maines TR, Jayaraman A, Belser JA, Wadford DA, Pappas C, et al. (2009) Transmission and Pathogenesis of Swine-Origin 2009 A(H1N1) Influenza Viruses in Ferrets and Mice. Science.

11. Reid AH, Janczewski TA, Lourens RM, Elliot AJ, Daniels RS, et al. (2003) 1918 influenza pandemic caused by highly conserved viruses with two receptor-binding variants. Emerg Infect Dis 9: 1249-1253.

12. Logan WP, Mac KD (1951) Development of influenza epidemics. Lancet 1: 264-265.

13. Collins SD, Lehmann J (1951) Trends and epidemics of influenza and pneumonia: 1918-1951. Public Health Rep 66: 1487-1516.

14. Rasmussen AF, Jr., Stokes JC, Smadel JE (1948) The Army experience with influenza, 1946-1947; laboratory aspects. Am J Hyg 47: 142-149.

15. Sartwell PE, Long AP (1948) The Army experience with influenza, 1946-1947; epidemiological aspects. Am J Hyg 47: 135-141.

16. Salk JE, Suriano PC (1949) Importance of antigenic composition of influenza virus vaccine in protecting against the natural disease; observations during the winter of 1947-1948. Am J Public Health Nations Health 39: 345-355.

17. Kilbourne ED, Smith C, Brett I, Pokorny BA, Johansson B, et al. (2002) The total influenza vaccine failure of 1947 revisited: major intrasubtypic antigenic change can explain failure of vaccine in a post-World War II epidemic. Proc Natl Acad Sci U S A 99: 10748-10752.

18. Isaacs A, Gledhill AW, Andrewes CH (1952) Influenza A viruses; laboratory studies, with special reference to European outbreak of 1950-1. Bull World Health Organ

6: 287-315.

19. Viboud C, Tam T, Fleming D, Miller MA, Simonsen L (2006) 1951 influenza epidemic, England and Wales, Canada, and the United States. Emerg Infect Dis 12: 661-668.

20. Viboud C, Tam T, Fleming D, Handel A, Miller MA, et al. (2006) Transmissibility and mortality impact of epidemic and pandemic influenza, with emphasis on the unusually deadly 1951 epidemic. Vaccine 24: 6701-6707.

21. Scholtissek C, Rohde W, Von Hoyningen V, Rott R (1978) On the origin of the human influenza virus subtypes H2N2 and H3N2. Virology 87: 13-20.

22. Nakajima K, Desselberger U, Palese P (1978) Recent human influenza A (H1N1) viruses are closely related genetically to strains isolated in 1950. Nature 274: 334-339.

23. Kendal AP, Noble GR, Skehel JJ, Dowdle WR (1978) Antigenic similarity of influenza A (H1N1) viruses from epidemics in 1977--1978 to "Scandinavian" strains isolated in epidemics of 1950--1951. Virology 89: 632-636.

24. Scholtissek C, von Hoyningen V, Rott R (1978) Genetic relatedness between the new 1977 epidemic strains (H1N1) of influenza and human influenza strains isolated between 1947 and 1957 (H1N1). Virology 89: 613-617.

25. Taubenberger JK, Reid AH, Janczewski TA, Fanning TG (2001) Integrating historical, clinical and molecular genetic data in order to explain the origin and virulence of the 1918 Spanish influenza virus. Philos Trans R Soc Lond B Biol Sci 356: 1829-1839.

26. Shope RE (1931) The Etiology of Swine Influenza. Science 73: 214-215.

27. Morens DM, Taubenberger JK, Fauci AS (2009) The Persistent Legacy of the 1918 Influenza Virus. N Engl J Med.

28. Brown IH (2000) The epidemiology and evolution of influenza viruses in pigs. Vet Microbiol 74: 29-46.

29. Olsen CW (2002) The emergence of novel swine influenza viruses in North America. Virus Res 85: 199-210.

30. Pensaert M, Ottis K, Vandeputte J, Kaplan MM, Bachmann PA (1981) Evidence for the natural transmission of influenza A virus from wild ducts to swine and its potential importance for man. Bull World Health Organ 59: 75-78.

31. Brown IH, Ludwig S, Olsen CW, Hannoun C, Scholtissek C, et al. (1997) Antigenic and genetic analyses of H1N1 influenza A viruses from European pigs. J Gen Virol 78 (Pt 3): 553-562.

32. Donatelli I, Campitelli L, Castrucci MR, Ruggieri A, Sidoli L, et al. (1991) Detection of two antigenic subpopulations of A(H1N1) influenza viruses from pigs: antigenic drift or interspecies transmission? J Med Virol 34: 248-257.

33. Reid AH, Fanning TG, Janczewski TA, Lourens RM, Taubenberger JK (2004) Novel origin of the 1918 pandemic influenza virus nucleoprotein gene. J Virol 78: 12462-12470.

34. Dunham EJ, Dugan VG, Kaser EK, Perkins SE, Brown IH, et al. (2009) Different evolutionary trajectories of European avian-like and classical swine H1N1 influenza A viruses. J Virol 83: 5485-5494.

35. Brown IH, Harris PA, McCauley JW, Alexander DJ (1998) Multiple genetic reassortment of avian and human influenza A viruses in European pigs, resulting

in the emergence of an H1N2 virus of novel genotype. J Gen Virol 79 (Pt 12): 2947-2955.

36. Webby RJ, Swenson SL, Krauss SL, Gerrish PJ, Goyal SM, et al. (2000) Evolution of swine H3N2 influenza viruses in the United States. J Virol 74: 8243-8251.

37. Newman AP, Reisdorf E, Beinemann J, Uyeki TM, Balish A, et al. (2008) Human case of swine influenza A (H1N1) triple reassortant virus infection, Wisconsin. Emerg Infect Dis 14: 1470-1472.

38. Shinde V, Bridges CB, Uyeki TM, Shu B, Balish A, et al. (2009) Triple-reassortant swine influenza A (H1) in humans in the United States, 2005-2009. N Engl J Med 360: 2616-2625.

39. Skehel JJ, Wiley DC (2000) Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu Rev Biochem 69: 531-569.

40. Tumpey TM, Maines TR, Van Hoeven N, Glaser L, Solorzano A, et al. (2007) A two-amino acid change in the hemagglutinin of the 1918 influenza virus abolishes transmission. Science 315: 655-659.

41. Srinivasan A, Viswanathan K, Raman R, Chandrasekaran A, Raguram S, et al. (2008) Quantitative biochemical rationale for differences in transmissibility of 1918 pandemic influenza A viruses. Proc Natl Acad Sci U S A 105: 2800-2805.

42. Stevens J, Blixt O, Glaser L, Taubenberger JK, Palese P, et al. (2006) Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities. J Mol Biol 355: 1143-1155.

43. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999) Predicting the evolution of human influenza A. Science 286: 1921-1925.

44. Bush RM, Fitch WM, Bender CA, Cox NJ (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol 16: 1457-1465.

45. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591.

46. Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46: 409-418.

47. Delport W, Scheffler K, Seoighe C (2008) Frequent toggling between alternative amino acids is driven by selection in HIV-1. PLoS Pathog 4: e1000242.

48. Seoighe C, Ketwaroo F, Pillay V, Scheffler K, Wood N, et al. (2007) A model of directional selection applied to the evolution of drug resistance in HIV-1. Mol Biol Evol 24: 1025-1031.

49. Kosakovsky Pond SL, Poon AF, Leigh Brown AJ, Frost SD (2008) A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. Mol Biol Evol 25: 1809-1824.

50. Anisimova M, Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol Biol Evol 26: 255-271.

51. Delport W, Scheffler K, Seoighe C (2009) Models of coding sequence evolution. Brief Bioinform 10: 97-109.

52. Lemey P, M. S, Vandamme A (2009) The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. Cambridge: Cambridge University Press.

53. Yang Z (2006) Computational Molecular Evolution. Oxford: Oxford University Press.

54. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for

heterogeneous selection pressure at amino acid sites. Genetics 155: 431-449.

55. Pond SL, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676-679.

56. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574.

57. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754-755.

58. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568-573.

59. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929-936.

60. Yang Z (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J Mol Evol 51: 423-432.

61. Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19: 908-917.

62. Robertson JS, Bootman JS, Newman R, Oxford JS, Daniels RS, et al. (1987) Structural changes in the haemagglutinin which accompany egg adaptation of an influenza A(H1N1) virus. Virology 160: 31-37.

63. Xu X, Rocha EP, Regenery HL, Kendal AP, Cox NJ (1993) Genetic and antigenic analyses of influenza A (H1N1) viruses, 1986-1991. Virus Res 28: 37-55.

64. Gambaryan AS, Robertson JS, Matrosovich MN (1999) Effects of egg-adaptation on the receptor-binding properties of human influenza A and B viruses. Virology 258: 232-239.

65. Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164: 1229-1236.

66. Heath L, van der Walt E, Varsani A, Martin DP (2006) Recombination patterns in aphthoviruses mirror those found in other picornaviruses. J Virol 80: 11827-11832.

67. Nelson MI, Holmes EC (2007) The evolution of epidemic influenza. Nat Rev Genet 8: 196-205.

68. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948.

69. (2009) Serum cross-reactive antibody response to a novel influenza A (H1N1) virus after vaccination with seasonal influenza vaccine. MMWR Morb Mortal Wkly Rep 58: 521-524.

70. Finkelman BS, Viboud C, Koelle K, Ferrari MJ, Bharti N, et al. (2007) Global patterns in seasonal activity of influenza A/H3N2, A/H1N1, and B from 1997 to 2005: viral coexistence and latitudinal gradients. PLoS One 2: e1296.

71. Wolf YI, Viboud C, Holmes EC, Koonin EV, Lipman DJ (2006) Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. Biol Direct 1: 34.

72. Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22: 1107-1118.

73. Caton AJ, Brownlee GG, Yewdell JW, Gerhard W (1982) The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell 31: 417-427.

74. Gerhard W, Yewdell J, Frankel ME, Webster R (1981) Antigenic structure of influenza virus haemagglutinin defined by hybridoma antibodies. Nature 290: 713-717.

75. Rogers GN, D'Souza BL (1989) Receptor binding properties of human and animal H1 influenza virus isolates. Virology 173: 317-322.

76. Matrosovich MN, Gambaryan AS, Teneberg S, Piskarev VE, Yamnikova SS, et al. (1997) Avian influenza A viruses differ from human viruses by recognition of sialyloligosaccharides and gangliosides and by a higher conservation of the HA receptor-binding site. Virology 233: 224-234.

77. Nobusawa E, Nakajima K, Nakajima S (1987) Determination of the epitope 264 on the hemagglutinin molecule of influenza H1N1 virus by site-specific mutagenesis. Virology 159: 10-19.

78. Gamblin SJ, Haire LF, Russell RJ, Stevens DJ, Xiao B, et al. (2004) The structure and receptor binding properties of the 1918 influenza hemagglutinin. Science 303: 1838-1842.

79. Gambaryan AS, Tuzikov AB, Piskarev VE, Yamnikova SS, Lvov DK, et al. (1997) Specification of receptor-binding phenotypes of influenza virus isolates from different hosts using synthetic sialylglycopolymers: non-egg-adapted human H1 and H3 influenza A and influenza B viruses share a common high binding affinity for 6'-sialyl(N-acetyllactosamine). Virology 232: 345-350.

80. Shen J, Kirk BD, Ma J, Wang Q (2009) Diversifying selective pressure on influenza B virus hemagglutinin. J Med Virol 81: 114-124.

81. Shi W, Gibbs MJ, Zhang Y, Zhuang D, Dun A, et al. (2008) The variable codons of H5N1 avian influenza A virus haemagglutinin genes. Sci China C Life Sci 51:

987-993.

82. Daniels RS, Douglas AR, Skehel JJ, Wiley DC, Naeve CW, et al. (1984) Antigenic analyses of influenza virus haemagglutinins with different receptor-binding specificities. Virology 138: 174-177.

83. Wiley DC, Wilson IA, Skehel JJ (1981) Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature 289: 373-378.

84. Sheerar MG, Easterday BC, Hinshaw VS (1989) Antigenic conservation of H1N1 swine influenza viruses. J Gen Virol 70 (Pt 12): 3297-3303.

85. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555-556.

86. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696-704.

87. Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12: 357-358.